LES ÉVALUATIONS STANDARDISÉES DES ÉLÈVES

Perspective historique

Bruno Trosseille et **Thierry Rocher**
MENESR-DEPP, bureau de l'évaluation des élèves

Depuis une quarantaine d'années, le ministère de l'Éducation nationale a mis en œuvre des évaluations tantôt « de masse », tantôt sur échantillons. Ces évaluations peuvent avoir deux fonctions principales : de diagnostic lorsqu'elles sont élaborées pour fournir aux enseignants des outils professionnels qui leur sont nécessaires pour adapter leur enseignement en fonction des acquis de leurs élèves ; de bilan lorsque l'objectif est d'observer les acquis des élèves et leur évolution pour le pilotage d'ensemble du système éducatif. La confusion, dans une même évaluation, de ces deux fonctions est potentiellement source d'erreurs et de troubles, tant sur le plan scientifique que sociétal. Après avoir décrit l'histoire entrelacée de ces deux types d'évaluations au sein du Ministère, nous envisageons l'avenir du paysage évaluatif et la façon dont il peut se réorganiser en fonction des différentes finalités qui lui sont aujourd'hui assignées et des défis qu'il devra affronter à l'avenir.

Depuis quatre décennies, la DEPP (et les services et directions qui l'ont précédée)[1] met en place des dispositifs d'évaluation, spécifiques, nationaux, des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'œuvre pour la France de diverses évaluations internationales (voir *infra*). Le développement des évaluations standardisées apparaît en effet, aux yeux des responsables des services statistiques, s'appuyant sur les exemples étrangers et internationaux, comme un complément indispensable des statistiques pour rendre compte du système et le piloter. Trois grandes périodes, qui se recouvrent peu ou prou, caractérisent le développement de ces dispositifs au ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche (MENESR).

1. On conviendra, par commodité, d'utiliser le sigle DEPP pour dénommer l'ensemble des services et directions qui ont précédé l'actuelle direction de l'évaluation, de la prospective et de la performance avec des missions d'évaluation (à savoir SEIS, SIGES, SPRESE, DEP, DPD).

Dans une première période, de la fin des années 1970 à la fin des années 1980, les dispositifs mis en œuvre à l'école et au collège, couvrent progressivement, niveau scolaire par niveau scolaire, l'ensemble des disciplines. Il s'agit, au regard des programmes en vigueur, d'établir un constat, d'apprécier l'état du système, de rendre compte des acquisitions des élèves aux responsables de la politique éducative du Ministère. Il s'agit aussi, déjà, de nourrir le débat public, ce qui sera plus nettement affiché au cours de la deuxième période. Dans le cadre d'un « observatoire permanent des acquis des élèves » des dispositifs d'évaluations sont systématiquement organisés, selon une méthodologie rigoureuse [Le Guen, 1991], sur des échantillons d'élèves, en fin d'année scolaire. Ils concernent la cinquième dès 1975, puis le CP dès 1979, et ensuite pratiquement tous les niveaux du primaire au lycée. Durant cette décennie, chaque niveau scolaire fera l'objet, une année donnée, d'une évaluation de type « bilan ».

Une deuxième période, à partir de 1989, est occupée par la mise en place d'évaluations diagnostiques « de masse », conséquences de la loi sur l'éducation de 1989, dite « loi Jospin ». Le rapport qui lui est annexé, relevant que *« moins d'un élève sur deux arrive au collège avec une maîtrise suffisante de la lecture »,* précise déjà l'urgence de la mise en œuvre d'un véritable plan sur l'apprentissage de la lecture et indique que *« cette acquisition fondamentale fera l'objet d'une évaluation auprès de tous les élèves entrant en cours élémentaire deuxième année et en sixième ; elle sera suivie d'actions de soutien ou de reprises d'apprentissage dans chaque école et chaque établissement scolaire »*. Dans son introduction à la revue *Éducation & formations* consacrée aux résultats nationaux des évaluations de septembre 1989, le ministre rappelle l'objectif de cette évaluation *« conçue comme un outil mis à votre disposition pour déceler, de façon précise et dès le début de l'année scolaire, les difficultés de vos élèves et vous permettre, dans toute la mesure du possible d'y apporter rapidement une réponse »* [*Éducation & formations,* 1990]. Les évaluations sur échantillons du type de celles de la première décennie se font alors plus rares (1990, 1995, 1999).

La troisième période, depuis le début du XXIe siècle, voit se systématiser, à côté des enquêtes internationales, notamment PISA[2] qui débute en 2000, des évaluations sur échantillons pour un bilan des acquis des élèves en fin d'école primaire et en fin de collège (Cedre[3]) et pour le calcul des indicateurs de maîtrise des compétences du socle commun destinés aux projets de lois de finances (LOLF) ▸ **Encadré**.
Durant cette période, de nouvelles évaluations « de masse » apparaissent brièvement, ayant pour but le repérage des élèves en difficulté vis-à-vis du socle commun (objectif de mise en place de remédiations ciblées en début de CE1, de 2005 à 2007, puis en début de CM2, en 2007), puis d'évaluations « bilans-diagnostics » (en fin de CE1 et en milieu de CM2 de 2009 à 2012) succédant à l'arrêt des évaluations « de masse » en CE2 et en sixième.
Avant de retracer de façon plus détaillée l'histoire, au MENESR, des deux grands types de dispositifs que sont les évaluations diagnostiques « de masse » et les évaluations de type bilan, faisons un petit détour méthodologique afin d'identifier ce qui les distingue.

2. Programme international pour le suivi des acquis des élèves, mené par l'OCDE.
3. Cedre : cycle des évaluations disciplinaires réalisées sur échantillons.

LE CYCLE DES ÉVALUATIONS DISCIPLINAIRES RÉALISÉES SUR ÉCHANTILLON (CEDRE)

Ce cycle d'évaluations établit des bilans nationaux des acquis des élèves en fin d'école et en fin de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires en référence aux programmes. La présentation des résultats permet de situer les performances des élèves sur des échelles de niveau allant de la maîtrise pratiquement complète de ces compétences à une maîtrise bien moins assurée, voire très faible, de celles-ci. Renouvelées tous les six ans (tous les cinq ans à partir de 2012), ces évaluations permettent de répondre à la question de l'évolution du « niveau des élèves » au fil du temps.

Le Calendrier

2003 – 2009 – 2015 : maîtrise de la langue française et compétences générales.
2004 – 2010 – 2016 : langues étrangères.
2005 : attitudes à l'égard de la vie en société (non repris par la suite).
2006 – 2012 – 2017 : histoire-géographie, éducation civique.
2007 – 2013 – 2018 : sciences expérimentales.
2008 – 2014 – 2019 : mathématiques.

Les élèves concernés

La population visée est celle des élèves de CM2 et de troisième générale des collèges publics et privés sous contrat de France métropolitaine. Pour les écoles, un échantillon représentatif est constitué au niveau national et tous les élèves de CM2 y passent les évaluations. Pour les collèges, des classes de troisième sont sélectionnées aléatoirement en vue d'une représentativité nationale. Pour chaque niveau scolaire, de 5 000 à 10 000 élèves, répartis dans plusieurs centaines de classes, sont évalués. Pour plus de détails, le lecteur est invité à consulter l'article de Garcia, Le Cam, Rocher [dans ce numéro, p. 101].

La comparabilité entre les différentes années d'évaluation

Afin de pouvoir comparer les résultats des enquêtes entre la première itération et les suivantes, une partie des items de la première est reprise à l'identique dans l'évaluation suivante (items dits « d'ancrage »). La mise en œuvre de modèles psychométriques adaptés – les modèles de réponse à l'item [Rocher, dans ce numéro, p. 37] – permet d'assurer la comparabilité entre les enquêtes successives et de mesurer l'évolution dans le temps de la distribution des niveaux de compétence des élèves.

Évaluations orientées élèves *vs* populations

Les évaluations standardisées des acquis des élèves offrent des perspectives diverses, selon que l'on s'intéresse aux élèves pris individuellement comme sujets de leurs apprentissages ou que l'on s'intéresse à eux comme éléments d'une population sur laquelle on cherchera à recueillir des informations destinées à éclairer le fonctionnement du système éducatif. Il est ainsi très important de clarifier, dès la phase initiale de la construction d'un dispositif, l'usage qui sera fait des données à recueillir. Au plan conceptuel, on peut distinguer les dispositifs « d'évaluation diagnostique » et les dispositifs « d'évaluation bilan » qui constituent deux types d'évaluations qui ne se substituent pas l'un à l'autre. Ils diffèrent dans leurs objectifs, dans les modalités de mise en œuvre, dans l'exploitation et l'utilisation des résultats.

Axées sur les élèves, outils professionnels pour les enseignants, les **évaluations diagnostiques** permettent d'établir un diagnostic individuel – évaluer les points forts et les points faibles (freins aux apprentissages), d'aider l'enseignant à définir les actions pédagogiques adaptées à la situation de chacun et à réguler la programmation des apprentissages. Au niveau local, utilisées par les enseignants dans la gestion pédagogique de leur classe, les évaluations diagnostiques, descripteurs des

réussites et échecs de chaque élève, aides à la constitution de groupes de besoin, supports pour la réflexion pédagogique, doivent permettre de cerner individuellement les compétences et les difficultés de chaque élève et d'orienter le travail de chaque élève et de la classe en fonction des résultats. Elles doivent aussi permettre de dégager des priorités de formation continue « de proximité ». En raison de leurs objectifs, de leur conception et de leur renouvellement annuel, ce type d'évaluations n'est pas adapté à la comparaison dans le temps. Toutefois, si les conditions de passation et de correction sont respectées, leurs résultats peuvent être comparés dans l'espace, entre les différentes classes d'une même école ou entre les différents collèges d'un département.

Axées sur des populations, les évaluations-bilans, (comme Cedre, PISA, les indicateurs de la LOLF) sont des outils pour le pilotage d'ensemble du système éducatif. Leur méthodologie de construction s'appuie sur les méthodes de la mesure en éducation et les modèles psychométriques [Laveault et Grégoire, 2002]. Elles concernent de larges échantillons représentatifs d'établissements, de classes et d'élèves. Organisées, le plus souvent, en fin de cycles, elles révèlent, en référence aux objectifs de la politique éducative, les objectifs atteints et ceux qui ne le sont pas. Ces évaluations doivent permettre d'agir au niveau national sur les programmes des disciplines, sur les organisations des enseignements, sur les contextes de l'enseignement, sur des populations caractérisées. Sous certaines conditions méthodologiques, leurs résultats peuvent être comparés dans le temps.

On notera que, depuis les années 1990, le Ministère a souvent entretenu la **confusion entre ces deux types d'évaluations**. Comme le pointent les inspecteurs généraux dans un rapport consacré à l'évaluation des acquis des élèves [IGEN-IGAENR, 2005, p. 15], *« les évaluations CE2/sixième ont eu, dès le début, une fonction ambiguë »*. Ainsi, au niveau national, la DEPP organise entre 1989 et 2007, sur des échantillons représentatifs d'élèves, une remontée des résultats sur les évaluations « de masse » en CE2 et en sixième. Ces résultats nationaux sont présentés comme des repères destinés à aider les enseignants à faire une analyse individuelle des freins rencontrés par leurs élèves dans les apprentissages. S'ils ne peuvent – au sens psychométrique [Dickes, Tournois *et alii,* 1994] – être comparés d'une année à l'autre, ils ont pu, ici ou là, être utilisés comme des indicateurs d'évolution des acquis des élèves, voire du système éducatif. La DEPP juge ainsi parfois nécessaire de rappeler que *« les évaluations nationales CE2 et sixième, tout comme celles d'entrée en seconde de lycée, n'ont de valeur qu'annuelle puisque les supports des évaluations et les objectifs évalués diffèrent chaque année. Aussi, ces résultats ne peuvent-ils en aucun cas être utilisés à des fins de comparaisons d'une année sur l'autre et détournés de leur objet pédagogique »*[4].
Cette confusion a encore été accentuée entre 2009 et 2012 avec la mise en place des évaluations « de masse » en fin de CE1 et en milieu de CM2, évaluations dont les objectifs ont été présentés de façon quelque peu « flottante », hésitant entre le diagnostic et le bilan pour finalement leur assigner ces deux objectifs simultanés.

4. Circulaire n° 2000-091 du 23 juin 2000, *Bulletin officiel,* n° 25 du 29 juin 2000.

Décrivons maintenant de manière plus détaillée la façon dont le Ministère a utilisé et déployé ces deux grands types d'évaluation au cours des trente dernières années. Nous retracerons tout d'abord l'histoire des grandes évaluations diagnostiques « de masse » qui, du fait de leur nature exhaustive, ont pendant une vingtaine d'années été, auprès des enseignants, la face visible du travail de la DEPP en matière d'évaluation. Nous montrerons ensuite pourquoi et comment s'est peu à peu installé dans le paysage un ensemble d'évaluations bilans standardisées sur échantillons dont l'objectif, outre celui de faire un état des lieux des acquis des élèves, est de donner à voir les évolutions de ces acquis, en permettant des comparaisons temporelles et internationales. Enfin, nous envisagerons ce que pourrait être un paysage renouvelé des évaluations au MENESR.

LES ÉVALUATIONS DIAGNOSTIQUES « DE MASSE »

Les évaluations en début de CE2, de sixième et de seconde

Les premières évaluations nationales « de masse » sont mises en place pour accompagner la loi d'orientation sur l'éducation dite « loi Jospin », du 10 juillet 1989 ; à la rentrée 1989 au début du cycle des approfondissements (CE2) et du collège (sixième), en français et en mathématiques. La DEPP pilote la conception et la mise en œuvre de ces évaluations dont les protocoles sont élaborés avec les corps d'inspection, des enseignants et chefs d'établissement, des représentants des directions pédagogiques, des chercheurs et des techniciens. Elles sont alors accompagnées d'un important effort de formation piloté par les directions pédagogiques du Ministère : formation de 400 formateurs de formateurs, conception et large diffusion de documents de « remédiation ». De plus, et c'est une nouveauté dans ce Ministère, la formation à l'utilisation de ces outils est rendue obligatoire pour tous les instituteurs de CE2 et les professeurs de français et de mathématiques de sixième.
Des outils informatisés de saisie et d'exploitation des résultats sont progressivement mis à disposition des enseignants (Casimir, puis J'ADE) ainsi que des statistiques détaillées, des dossiers de synthèse et de commentaires approfondis, d'abord sur support papier, jusqu'en 2002 [Brézillon, Chollet-Remvikos *et alii,* 2003], puis en ligne jusqu'en 2007. Pour Le Guen, responsable du département de l'évaluation des élèves et des étudiants au sein de la DEPP, *« cinq types d'actions relèvent de cette évaluation-diagnostique »* [Le Guen, 1991]. Si les trois premières actions sont clairement orientées élèves et enseignants : aider à mieux connaître les lacunes individuelles des élèves ; fournir une méthodologie et des outils d'évaluation pour les apprentissages de base ; contribuer à une meilleure efficacité en étant un outil de dialogue avec les familles, les deux autres sont « axées populations » : fournir des indicateurs pour le pilotage du système ; rendre compte au niveau national de l'efficacité de l'école.

Huit ans après leur mise en place, analysant l'usage qui en est fait, Levasseur [1996] leur attribue cinq fonctions : outil professionnel, outil de dialogue avec les parents, outil d'adaptation de l'enseignement, outil de formation des enseignants, outil de régulation du système éducatif. Parallèlement, la DEPP propose aux enseignants, de la maternelle au lycée, dans toutes les disciplines et durant toute l'année scolaire,

une « banque d'outils d'aide à l'évaluation ». Son objectif est de donner aux enseignants des outils diversifiés pour analyser les compétences des élèves et de *« leur permettre de faire évoluer les progressions pédagogiques en fonction des besoins objectivement repérés chez les élèves de la classe »*. D'abord sur support papier, cette banque est informatisée et consultable en ligne dès 2002. Elle n'est cependant pas actualisée depuis 2006. Le coût d'élaboration d'une telle banque est en effet très élevé, surtout si l'on souhaite accompagner les outils de résultats statistiques fiables permettant de donner des repères aux utilisateurs.

À la rentrée 1992, de nouvelles évaluations diagnostiques sont proposées au début du lycée, dans quatre disciplines : français, mathématiques, histoire-géographie, première langue vivante pour les secondes générales et technologiques ; français, mathématiques, sciences et techniques industrielles, économie-gestion pour les secondes professionnelles. L'évaluation de tous les élèves à l'entrée en seconde, évaluation de compétences à visée diagnostique, est destinée à faciliter la mise en œuvre des modules et de l'aide individualisée (à partir de 1999) afin de répondre au mieux aux besoins des élèves dans leur diversité. Ayant perduré pendant presque dix années, ces évaluations en classe de seconde ne seront pas reconduites à la rentrée 2002, compte tenu de leur faible usage par les enseignants. En effet, en 2001, Claude Pair, ancien recteur d'académie, examinant à la demande du Haut Conseil de l'évaluation de l'école (HCéé) les « forces et faiblesses de l'évaluation du système éducatif » écrit :
« Il faut distinguer les niveaux CE2 et sixième de celui de seconde. Dans le premier cas, l'opération est acceptée et effectuée quasiment partout ; les résultats sont restitués aux parents. Comme cela nous a été dit, "elle est entrée dans le paysage" [...] *Mais en seconde la situation semble beaucoup moins favorable : d'après les avis recueillis, l'évaluation est loin d'être effectuée partout et elle est exploitée moins souvent encore, les résultats n'étant donc guère communiqués aux collèges dont viennent les élèves.* [...] *En outre, les enseignants de lycée sont peut-être moins réceptifs que ceux de l'école ou du collège au questionnement pédagogique créé par l'hétérogénéité des acquis des élèves. »* [Pair, 2001]
Dans un contexte, déjà, de ressources contraintes, la DEPP doit faire des choix et la mise en place, dès le début des années 2000 (suite à l'avis n° 2 du HCéé, juin 2001), des évaluations bilans, plus tard appelées Cedre, l'engagement dans les évaluations triennales internationales PISA et la construction d'indicateurs nouveaux dans le cadre de la LOLF (voir *infra*), nécessitent un redéploiement de ressources humaines et financières qui décident du sort des évaluations à l'entrée en seconde, jusqu'alors mises en œuvre par la DEPP. Une incursion en cinquième est cependant tentée à la rentrée scolaire 2002. Elle ne sera pas poursuivie.

Alors que le Ministère, souhaitant mettre davantage l'accent sur la prévention de l'illettrisme, prépare dès 2004 de nouvelles évaluations en début de CE1, l'idée s'impose progressivement d'arrêter les évaluations en CE2 et en sixième. Cette décision sera actée en janvier 2007, la circulaire de rentrée[5] indiquant que les évaluations sont supprimées en CE2 et reconduites pour la dernière fois en sixième à la rentrée 2007. Elles sont cependant reconduites en sixième à la rentrée 2008, pour satisfaire

5. Circulaire n° 2007-011 du 9 janvier 2007, *Bulletin officiel*, n° 3 du 18 janvier 2007.

la demande de l'encadrement de terrain qui souhaite leur maintien pour le pilotage pédagogique local. Toutefois, les groupes de concepteurs élaborant les protocoles d'évaluation (cahiers des élèves, consignes de passation et de correction pour les enseignants) ne sont pas reconduits et les protocoles utilisés restent identiques de 2005 à 2008, sans que les résultats relevés sur des échantillons représentatifs varient au cours des trois ou quatre dernières années de leur utilisation.

Pour répondre à la demande de repérage des élèves rencontrant le plus de difficultés, sont donc expérimentées[6] en 2004 et en 2005, et généralisées dès la rentrée 2006, de nouvelles évaluations en début de CE1, en lien avec les programmes personnalisés de réussite éducative (PPRE). Elles s'inscrivent dans le dispositif de prévention de l'illettrisme et en cohérence avec les nouveaux programmes de 2002. Elles seront suivies du même type d'évaluations en début de CM2, sous statut expérimental à la rentrée 2007. Ces évaluations comportent un premier filtre destiné à repérer les élèves (estimés en moyenne à 20 % de la population) nécessitant un regard plus fin sur leurs besoins. Une fois ces élèves repérés, il est demandé aux enseignants de leur administrer un deuxième livret d'évaluation permettant de cerner avec plus de précision leurs difficultés et de les orienter vers les remédiations les plus adaptées, à l'intérieur ou à l'extérieur de la classe. Le cahier des charges de l'évaluation en CE1 précise qu'elle ne fera pas l'objet de statistiques au niveau national et qu'il n'y aura donc pas de remontées vers le Ministère.
Cependant, dès 2006, dans une note interne du 16 octobre, le directeur de cabinet demande de produire pour le 20 décembre *« une synthèse des résultats de l'évaluation nationale de CE1 »* et, pour l'année scolaire suivante, la préparation d'un dispositif fonctionnel de remontée exhaustive des résultats des évaluations aux différents niveaux territoriaux pour *« permettre aux différents échelons de pilotage pédagogique du premier degré d'appréhender finement la situation réelle de nos élèves en matière de maîtrise des compétences du socle commun »*. Il est également demandé de prévoir, sur échantillon représentatif, une synthèse au niveau national pour l'information du ministre et de l'administration centrale.
Ce mélange des genres est mal reçu au niveau local comme en témoigne le rapport établi en mars 2007 par Claus et Mégard [IGEN, 2007, p. 9] sur le suivi de cette évaluation, qui en souligne les ambiguïtés en posant la question de la capacité d'une évaluation diagnostique à devenir un outil de pilotage : *« De manière générale, une évaluation dont le seul objectif est d'aider les enseignants à gérer l'hétérogénéité des élèves peut s'accommoder de différences lors de la passation des épreuves (consignes répétées ou expliquées, temps supplémentaire accordé, etc.). Une évaluation qui aboutit à des moyennes et des comparaisons pour piloter à une échelle supérieure à celle de l'école ne peut se satisfaire de telles approximations. Le changement de cap imposé en 2006 a eu pour conséquence la construction d'indicateurs peu fiables et donc peu exploitables. »*

Les inspecteurs recommandent fortement une clarification des finalités de l'évaluation des élèves au début du CE1, d'autant plus, ajoutent-ils, que cette évaluation sera accompagnée d'une évaluation de même nature au début du CM2. Ils ajoutent :

6. Circulaire n° 2004-015 du 25 janvier 2004, *Bulletin officiel,* n° 6 du 5 février 2004 et circulaire n° 2004-108 du 05 juillet 2004, parue au *Bulletin officiel* n° 28 du 15 juillet 2004.

« En tout état de cause, le Ministère doit impérativement faire preuve de constance. Une décision arrêtée et communiquée aux académies et aux écoles ne peut sans risque de rupture de confiance être abrogée sans raison impérative. Si les résultats d'une évaluation doivent être communiqués, publiés et agglomérés, il convient d'en informer les écoles avant la passation. »

À la rentrée 2007, les élèves de CE1 passent des protocoles modifiés et ceux de CM2 expérimentent un nouveau protocole. La circulaire de rentrée 2007 (voir note 4) précise la vocation essentiellement analytique de ces deux évaluations qui *« visent à repérer et analyser les difficultés et les freins que rencontrent certains élèves de CE1 et de CM2 dans leurs apprentissages dans les domaines de la lecture, de l'écriture et des mathématiques pour acquérir les compétences du socle attendues en fin des paliers 1 et 2* [...]*. Ces évaluations s'insèrent tout naturellement dans le dispositif mis en place localement pour déterminer si un élève doit bénéficier d'un programme personnalisé de réussite éducative (PPRE) et pour envisager les modalités de celui-ci. »*
Bien que soit annoncé le calcul de scores nationaux à partir d'un échantillon représentatif[7], aucun bilan de ces évaluations ne sera finalement réalisé. Elles disparaîtront sans bruit du paysage dès la fin de l'automne 2007 avec la préparation de nouvelles évaluations nationales obligatoires et exhaustives en CE1 et en CM2, qui auront cours de 2009 à 2012 et dont le pilotage sera confié à la DGESCO. Ce transfert d'un pilotage qui n'entre pas dans les missions habituellement dévolues à cette direction peut être compris comme la volonté de donner à cette dernière les moyens de contrôle dont elle souhaite disposer pour s'assurer que les réformes pédagogiques qu'elle impulse sont bien mises en œuvre au niveau local. Les nouveaux programmes de l'école primaire applicables à partir de la rentrée 2008 en sont notamment l'enjeu et, ces nouvelles évaluations, l'instrument. C'est d'ailleurs ce que notent les inspecteurs généraux Claus et Roze dans leur troisième note de synthèse à destination du ministre : « *Ces évaluations révèlent aussi l'écart, qui peut être important, entre ce qui est enseigné et ce qui devrait l'être. En ce sens les évaluations nationales sont un puissant levier pour une mise en œuvre complète des nouveaux programmes.* » [IGEN-IGAENR, 2009, p. 10].

Les évaluations en CE1 et en CM2

Le 5 juillet 2007, la lettre de mission du président de la République à son ministre de l'éducation[8] souhaite l'organisation d'*« une évaluation systématique de tous les élèves tous les ans, afin de repérer immédiatement les élèves en difficulté et de pouvoir les aider ; une évaluation régulière des enseignants sur la base des progrès et des résultats de leurs élèves »*. Ces nouvelles évaluations en CE1-CM2 sont en totale rupture avec les évaluations précédentes à ces niveaux scolaires. Situées en fin d'année scolaire pour le CE1, en janvier pour le CM2, elles sont présentées au départ comme des bilans devant également servir à évaluer les enseignants. La publication des résultats sur Internet, école par école, est même annoncée. Sollicitée par le Cabinet, la DEPP fait part de ses réserves et soulève un certain nombre d'interrogations, notamment sur les usages de l'évaluation, la comparabilité temporelle et la prise

7. Circulaire n° 2007-140 du 23 août 2007, *Bulletin officiel,* n° 30 du 30 août 2007.
8. http://discours.vie-publique.fr/notices/077002457.html

en compte du contexte social de l'école. Devant la levée de boucliers suscitée tant chez les enseignants que chez les parents d'élèves, l'idée de la publication des résultats école par école fait long feu. Toutefois, subsiste chez les enseignants une défiance quant à la vraie nature de ces évaluations, présentées à la fois comme bilan et comme diagnostic, en insistant tantôt sur un aspect, tantôt sur l'autre, et pouvant servir à contrôler leur valeur professionnelle. Cet usage possible de l'évaluation est ressenti comme d'autant plus injuste qu'il ne repose pas sur les progrès réalisés par les élèves, mais uniquement sur leur niveau à un instant donné, sans prendre en considération leur niveau scolaire à leur arrivée dans la classe ni leurs différences socioéconomiques. Cette confusion amène une résistance jamais encore vue chez les enseignants du primaire contre des évaluations malgré une prime de 400 € instituée pour les enseignants des niveaux concernés.

Dans un document d'orientation de novembre 2007[9], ces évaluations sont affichées comme une innovation : *« Il faut également se donner les moyens de connaître et de faire connaître quels sont les acquis des écoliers français à des moments clefs de leur scolarité, notamment par rapport aux pays comparables. C'est pourquoi seront créées* ***deux évaluations nationales témoins qui serviront à mesurer les acquis des élèves au CE1 et au CM2****.* [...] *Leurs constats seront rendus publics et permettront d'apprécier l'évolution de la réussite du système éducatif. »* Cette présentation ne fait aucune référence à l'existence d'évaluations spécifiquement construites pour assurer des comparaisons temporelles ou internationales, telles que les enquêtes nationales du cycle Cedre (voir *infra*), basées sur les programmes et mises en place depuis 2003, et l'enquête internationale PIRLS[10] qui existe depuis 2001.

En complément, la circulaire de rentrée 2008[11] précise que ces nouveaux protocoles d'évaluation *« permettent de dresser un bilan des acquis des élèves en CE1 et en CM2, premiers paliers du socle commun.* [...] *Les résultats scolaires des élèves seront un élément essentiel du pilotage. »* Interrogé sur ces évaluations par le site « Le café pédagogique »[12], le chef du bureau des écoles de la DGESCO déclare : *« Ce sont des évaluations organisées autour des programmes. C'est la grande différence avec les évaluations précédentes. La référence c'est le programme. Par conséquent on a affaire à une évaluation bilan de ce que les élèves ont acquis. En même temps, quand on regarde ce qui n'a pas été réussi on est sur le versant du repérage voire du diagnostic. »* On le voit, pour ces évaluations, la double assignation de bilan et de diagnostic est clairement assumée. Elle sera critiquée tant par les organisations syndicales que par le Haut Conseil de l'Éducation [voir *infra*, HCÉ, 2011] ou encore dans le rapport du groupe UMP de l'Assemblée nationale [Breton et Marc, 2009] qui souligne dans sa conclusion : *« **Une clarification des objectifs poursuivis est nécessaire**. En effet, les évaluations mises en place pour les élèves de CM2 sont à mi-chemin entre l'évaluation-bilan et l'évaluation-diagnostic appelant un dispositif de remédiation. Cette question ne semble pas complètement tranchée et une clarification par le ministère de l'Éducation nationale, en concertation avec l'ensemble des acteurs concernés, serait utile sinon indispensable. »*

9. Document d'orientation. Propositions du ministre de l'Éducation nationale, soumises à discussion, pour définir un nouvel horizon pour l'école primaire : http://media.education.gouv.fr/file/40/9/20409.pdf.
10. Le programme international sur la recherche en lecture scolaire, mené par l'IEA (*International Association for the Evaluation of Educational Achievement*) évalue les compétences en lecture des élèves en quatrième année de scolarité obligatoire (CM1 pour la France).
11. Circulaire n° 2008-042 du 4 avril 2008, *Bulletin officiel,* n° 15 du 10 avril 2008.
12. http://www.cafepedagogique.net/lesdossiers/pages/2009/evacm2ministere.aspx

En termes de fiabilité, une étude interne, réalisée par la DEPP lors de la première évaluation de janvier 2009, fait apparaître des distorsions dans les résultats selon que les écoles ont ou non été suivies par les inspecteurs du contrôle qualité, ainsi qu'en fonction des secteurs de scolarisation[13]. Ainsi, on observe une surestimation des élèves par leurs enseignants, et ce de façon plus particulièrement marquée dans le secteur privé, en l'absence de contrôle des procédures de passation et de correction. Dès la deuxième année d'utilisation, les limites de l'exercice, en termes de comparabilité, sont atteintes : les résultats des élèves de CM2 affichent une forte baisse en mathématiques. Cette baisse est en fait due à la plus grande difficulté du protocole élaboré pour cette deuxième itération, mais elle est interprétée comme une perte de compétence moyenne des élèves de CM2. La DEPP, sollicitée pour donner une mesure de l'impact de cette absence de contrôle de l'élaboration des protocoles, utilise une procédure d'*equating* (mise à niveau des métriques) pour permettre la comparabilité entre les deux années [Rocher, 2012]. Mais la suspicion à l'égard de ces évaluations est telle que l'ajustement des résultats de cette deuxième évaluation est dénoncé par beaucoup comme un « bidouillage » destiné à masquer l'impéritie du Ministère.

Celles-ci seront menées durant quatre années (de janvier 2009 à juin 2012) et ne seront pas reconduites après le changement de gouvernement de mai 2012. Le ministre Vincent Peillon indique qu'elles pourront être utilisées localement mais décide l'arrêt des « remontées » des informations à l'administration centrale, puisque « *les outils qui sont actuellement utilisés ne permettent pas une évaluation scientifiquement incontestable du système éducatif national* »[14]. Leur utilisation sera rendue facultative en 2013[15], puis abandonnée en 2014 (voir *infra*).

LES ÉVALUATIONS « BILANS » STANDARDISÉES À GRANDE ÉCHELLE

Présentes au Ministère dès la fin des années 1970, les enquêtes portant sur l'évaluation des compétences, appelées aussi évaluations standardisées à grande échelle, se sont multipliées depuis le début des années 2000 (voir plus loin la présentation de ces dispositifs). Leur objectif principal est de rendre compte de résultats au-delà du niveau individuel, en l'occurrence au niveau national et international. Le fait de leur assigner un objectif de représentativité implique des contraintes spécifiques dans leur élaboration.

Ce type d'évaluations occupe une place importante dans le débat sur l'éducation, notamment *via* la médiatisation – voire l'instrumentalisation politique – de leurs résultats [Mons, 2008]. La mise en œuvre de politiques éducatives se réfère aujourd'hui systématiquement à ces évaluations, en particulier aux évaluations internationales qui, derrière la diffusion de palmarès globalisants, fournissent un éclairage important sur les forces et les faiblesses des systèmes éducatifs [voir par exemple : Mons, 2007 ; Rocher, 2008b ; Baudelot et Establet, 2009 ; Rocher et Le Donné, 2012a].

13. Rapport conjoint IGEN-IGAENR : deuxième note de synthèse sur l'évaluation des élèves de CM2, n° 2009-028 du 31 mars 2009 et note interne DEPP non publiée.
14. Communiqué de presse du 21 mai 2012.
15. Circulaire n° 2013-060 du 10 avril 2013, *Bulletin officiel,* n° 15 du 11 avril 2013.

La DEPP, consciente de l'importance politique et scientifique de ces évaluations, obtient dès la décennie 1990 d'en être l'opérateur en France pour le compte des institutions qui les organisent (OCDE, IEA, Union européenne)[16].

Ces évaluations-bilans vouent une attention particulière aux comparaisons temporelles, afin de pouvoir juger des progrès réalisés par les systèmes éducatifs et de les rapprocher de caractéristiques structurelles, sociales, éducatives, etc. La question de la mesure de l'évolution du niveau des élèves dans le temps est donc centrale. Pourtant, dans l'histoire de l'évaluation et des tests, les études qui visent à comparer les compétences des sujets à différentes époques sont relativement rares.

Dans le domaine de l'intelligence cependant, des études comparatives assez anciennes ont révélé le fameux « effet Flynn », c'est-à-dire l'élévation des performances à des tests d'intelligence [Flieller, 2001]. Comme le notent Flieller, Manciaux et Kop [1995], ces enquêtes méritent qu'on leur prête attention pour deux raisons majeures. La première est d'ordre théorique : il s'agit de se prononcer sur le caractère absolu de certaines lois psychologiques, en appréciant leur permanence à travers différentes périodes de l'histoire. La seconde est d'ordre « pratique » : ces enquêtes permettent de répondre de manière objective à la demande sociale récurrente qui concerne l'évolution du niveau d'intelligence, de connaissances ou de compétences de la population. Pour répondre à cette nécessaire objectivation, ces enquêtes s'appuient sur des principes méthodologiques bien établis [Rocher, dans ce numéro, p. 37].

En France, l'intérêt pour ces évaluations à visée comparative trouve son origine dans le débat sur la baisse supposée du niveau scolaire des élèves, débat qui semble être particulièrement vif en France à la fin des années 1980. En 1992, dans un rapport au ministre de l'Éducation nationale, Claude Thélot, alors directeur de la DEPP, s'interroge sur les raisons de ce sentiment d'inquiétude et dégage trois pistes d'explication [Thélot, 1992]. Premièrement, après une période de massification du système éducatif, l'attention est portée sur la qualité et donc le niveau de compétence des élèves. Un système « de masse » peut-il être performant ? Deuxièmement, dans la lignée du rapport alarmiste américain *A Nation at Risk* de 1983 [National Commission on Excellence in Education, 1983], avec le renforcement de la compétition économique au niveau international, l'élévation du niveau des élèves apparaît comme un levier indispensable. Enfin, le sentiment de déclin du système éducatif porterait moins sur les mathématiques et les sciences, disciplines valorisées socialement, que sur les lettres et les humanités. Selon Thélot [1992], ce distinguo traduirait une appréhension profonde quant à l'avenir du pays, de sa langue et de son identité.

Néanmoins, Thélot [1992], tout comme Baudelot et Establet [1989], dans leur ouvrage *Le niveau monte*, soulignent le manque de mesures directes et objectives de l'évolution des acquis des élèves, alors même que cette question fait

16. C'est également dans cette perspective qu'à l'initiative de la DEPP, se met en place en 1997 le « Consortium université de Bourgogne », pour répondre à l'appel d'offres lancé par l'OCDE pour la mise au point et l'organisation de l'enquête PISA 2000 [Bottani et Vrignaud, 2005, p. 159].

l'objet d'une forte demande politique et sociale[17]. À l'époque, les seules données disponibles permettant une comparaison temporelle moins subjective sont celles issues des tests passés pendant les « trois jours » organisés par le ministère de la Défense [Baudelot et Establet, 1988]. Même si ces évaluations ne concernaient que les garçons, elles donnaient une bonne approximation de l'évolution du « niveau », ne serait-ce qu'en raison du nombre important de ceux qui les passaient et de leur proximité avec une génération entière de garçons.

Forte de ce constat, dans les années 1990, la DEPP conduit plusieurs études visant à mesurer l'évolution des acquis des élèves, comme la comparaison des compétences en français et en calcul des élèves des années 1920 à celles des élèves de 1995 qui avait clairement pour objectif de répondre aux tenants de la faillite du système éducatif, souvent nostalgiques d'un modèle scolaire révolu [Dejonghe, Levasseur *et alii*, 1996 ; Pons, 1996]. D'autres enquêtes ont concerné les élèves de troisième [Dessus, Jouvanceau, Murat, 1996] ou les élèves les plus performants scolairement [Peretti, Petrone, Thélot, 1996].

Cependant, les méthodes psychométriques permettant de construire des comparaisons diachroniques fiables apparaissent alors insuffisamment connues à la DEPP et plus généralement, à l'Insee ou dans l'enseignement des statistiques en France. Pourtant, des travaux reposant sur une méthodologie psychométrique adaptée à la comparaison diachronique existent à cette époque, dans le champ de la psychologie différentielle, comme en témoignent les enquêtes sur le niveau intellectuel des jeunes enfants [Flieller, Santigny, Schaeffer, 1986 ; Flieller, Manciaux, Kop, 1995].
Les premières enquêtes comparatives menées par la DEPP montrent, quant à elles, quelques faiblesses méthodologiques. Le recours à l'expertise de l'Inetop (Institut national d'étude du travail et d'orientation professionnelle) sur la comparabilité d'évaluations ayant eu lieu en sixième [Bonora et Vrignaud, 1996] et en troisième [Bonora et Vrignaud, 1997] permet de pointer les problèmes de comparabilité. Ces rapports sont aussi l'occasion d'introduire à la DEPP une connaissance des modèles de réponse à l'item (MRI), suite aux polémiques autour de l'enquête IALS[18] [Blum et Guérin-Pace, 2000 ; Murat et Rocher, dans ce numéro, p. 83]. En 1997, la reprise de l'évaluation LEC (lire, écrire, compter) de 1987 fait alors l'objet d'analyses psychométriques appropriées.
Comme on l'a vu plus haut, la France a pourtant une longue expérience des évaluations standardisées des élèves, à travers la mise en place des évaluations nationales diagnostiques de CE2 et de sixième. Malheureusement, aucun ajustement de la difficulté des épreuves n'a été entrepris afin de distinguer ce qui relève de la difficulté des épreuves de ce qui relève du niveau des élèves. En effet, nous l'avons rappelé, l'objectif premier de ces évaluations n'était pas de rendre compte de l'évolution du niveau des élèves dans le temps, mais de servir d'outils de repérage des difficultés pour les enseignants.

17. Paradoxalement, les évaluations des élèves sont très présentes dans le système scolaire français, à travers les contrôles continus fréquents conduits par les enseignants. Des études docimologiques, menées depuis près d'un siècle, notamment dans le cadre des travaux de la commission Carnegie sur le baccalauréat en 1936, montrent pourtant que le jugement des élèves par les enseignants est en partie empreint de subjectivité et peut dépendre de facteurs étrangers au niveau de compétence des élèves [voir par exemple : Piéron, 1963]. La notation des élèves est ainsi susceptible de varier sensiblement selon les caractéristiques des enseignants, des contextes scolaires, ainsi que des élèves eux-mêmes. Ces observations se retrouvent également aujourd'hui dans l'analyse des attestations de maîtrise des compétences du « socle commun de connaissances et de compétences » [Daussin, Rocher, Trosseille, 2010].
18. *International Adult Literacy Survey.*

L'essor des évaluations à visée diachronique

En 2001, l'avis du HCéé n° 2 [HCéé, 2001] pointe à son tour le manque d'informations objectives sur ce sujet et recommande la mise en place d'un dispositif *ad hoc* de suivi de l'évolution des acquis des élèves dans le temps [Salines et Vrignaud, 2001], comme cela existe dans d'autres pays, par exemple aux États-Unis où le dispositif NAEP (*National Assessment for Educational Progress*) fournit des séries de résultats comparables depuis la fin des années 1960 [Zwick, 1992 ; Jones et Olkin, 2004].
À la suite des recommandations du rapport de Salines et Vrignaud [2001], la DEPP donne naissance en 2003 au cycle des évaluations Cedre qui évalue les acquis des élèves de CM2 et de troisième, au regard de ce qui est attendu par les programmes scolaires. Chaque année, le domaine évalué est différent et à partir de 2009, des comparaisons temporelles concernent la maîtrise de la langue française [Colmant, Daussin, Bessonneau, 2011 ; Bourny, Bessonneau *et alii,* 2010), les langues étrangères [Bessonneau, Beuzon, Boucé *et alii.* 2012 ; Bessonneau, Beuzon, Daussin *et alii*, 2012], l'histoire-géographie [Garcia et Pastor, 2013 ; Garcia et Krop, 2013] et les sciences [Andreu, Étève, Garcia, 2014 ; Bret, Garcia, Roussel, 2014].
Depuis, d'autres dispositifs d'évaluations construits pour permettre des comparaisons diachroniques se sont développés en France ▸ **Tableau 1 p. 28-29**.

Plusieurs phénomènes relativement récents expliquent l'essor important de ces évaluations et leur multiplicité actuelle. Tout d'abord, au-delà de la demande du HCéé [Haut Conseil de l'évaluation de l'école, 2003], le souci de construire des indicateurs de suivi, pour le pilotage du système, est devenu de plus en plus prégnant, notamment dans le cadre de la LOLF, qui implique la construction d'indicateurs annuels de résultats [Rocher, 2008c]. Parallèlement, les évaluations internationales, telles que PISA [OCDE, 2013], PIRLS [Mullis, Martin *et alii,* 2012 ; Colmant et Le Cam, 2012] et TIMSS[19] [Mullis, Martin *et alii,* 2012], ont largement contribué à l'importance accordée aux comparaisons temporelles, en organisant les enquêtes de manière cyclique (voir tableau 1).

Alors que ces évaluations développées récemment ont été conçues de manière à assurer la comparaison diachronique, certains dispositifs proposent des comparaisons *ex post* de plus long terme, comme l'enquête SPEC6 [Rocher et Le Donné, 2012b] et l'enquête LEC [Rocher, 2008a]. Pour des synthèses concernant l'évolution des acquis des élèves, on se reportera à Rocher [2010] et à Daussin, Keskpaik, Rocher [2011].
Les évaluations réalisées dans le cadre des suivis longitudinaux de la DEPP (panels) permettent aussi de procéder à des comparaisons diachroniques, bien que ce ne soit pas leur objet premier qui est de décrire et d'expliquer les carrières et performances scolaires des élèves en rapprochant parcours scolaires, résultats aux évaluations et éléments de contexte. C'est le cas de la comparaison des acquis des élèves en début de CP de 1997 à 2011 [Le Cam, Rocher, Verlet, 2013].

Enfin, certains dispositifs concernent des sujets plus âgés. Ainsi, l'évaluation de la lecture, conçue par la DEPP et passée par tous les jeunes d'environ 17 ans lors de la JDC (Journée défense et citoyenneté), produit des indicateurs annuels de

19. *Trends in International Mathematics and Science Study.*

Tableau 1 Dispositifs d'évaluations standardisées en France permettant des comparaisons diachroniques

Test	Nom	Années
Évaluations nationales		
Cedre	Cycle des évaluations disciplinaires réalisées sur échantillons	Annuel, depuis 2003
LOLF	Évaluations pour les indicateurs de la LOLF	Annuel, depuis 2007
LEC	Lire, écrire, compter	1987, 1997, 2007
SPEC6	Étude spécifique des difficultés de lecture	1997, 2007
Panel CP	Évaluations standardisées des acquis des élèves du panel CP	1997, 2011
JDC	Journée défense et citoyenneté	Annuel, depuis 1998
IVQ	Information et vie quotidienne	2004, 2011
Évaluations internationales		
ESLC	*European Survey on Language Competences*	2011
PIRLS	*Progress in International Reading Literacy Study*	2001, 2006, 2011, 2016
PISA	*Programme for International Student Assessment*	Tous les trois ans depuis 2000
TIMSS	*Trends in International Mathematics and Science Study*	1995, 2015
Piaac	*Programme for the International Assessment of Adult Competencies*	2012

suivi de performance [voir par exemple, Vourc'h, Rivière *et alii,* 2014]. L'enquête IVQ (Information et vie quotidienne) évalue quant à elle un large échantillon d'adultes, et a permis une comparaison entre 2004 et 2011 [Jonas, 2012].

PERSPECTIVES

Comme nous l'avons indiqué en fin de partie 1, l'arrêt porté aux évaluations de CE1 et de CM2 était motivé par la clarification des objectifs des évaluations et par le rétablissement de la confiance du monde enseignant. Dès le printemps 2012, alors que se met en place la concertation pour la « Refondation de l'École », le ministre rétablit la distinction entre évaluations diagnostiques, outils professionnels des enseignants dans le sens mentionné plus haut, et évaluations bilans standardisées dont l'objectif est l'évaluation des acquis des élèves, indicateurs participant à l'évaluation du système éducatif[20].

20. Communiqué de presse du 21 mai 2012 et discours de Vincent Peillon lors de la conférence de presse de rentrée du 29 août 2012.

Population	Domaines
CM2 et 3e	Maîtrise de la langue (MDL) en CM2, compétences générales (CG) en troisième ; langues vivantes ; attitudes à l'égard de la vie en société ; histoire, géographie et éducation civique ; sciences expérimentales ; mathématiques.
CE1, CM2 et 3e	Compétences de base en français et en mathématiques, compétences du socle commun.
CM2	Compréhension de l'écrit, orthographe, calcul.
Début sixième	Automatismes, lexique, compréhension.
Début CP	Pré-lecture, écriture, numération, compréhension orale.
Environ 17 ans	Compréhension de l'écrit, lexique, automatismes.
16-65 ans	Littératie, numératie.
Troisième	Anglais, espagnol (compréhension de l'écrit et de l'oral).
CM1	Compréhension de l'écrit.
15 ans révolus	Compréhension de l'écrit, culture mathématique, culture scientifique.
CM1, TS	Mathématiques et sciences physiques.
16-65 ans	Littératie et numératie.

Aujourd'hui, d'une façon générale, l'évaluation des acquis des élèves peut répondre à trois objectifs :
– fournir aux enseignants des outils afin d'enrichir leurs pratiques pédagogiques en évaluant mieux les acquis de leurs élèves ;
– disposer d'indicateurs permettant de mesurer, au niveau national, les performances de notre système (évolutions temporelles et comparaisons internationales) ;
– doter les « pilotes de proximité » (recteurs, DASEN, IEN) d'indicateurs leur permettant de mieux connaître les résultats des écoles et d'effectuer une vraie régulation.
Quelle que soit la façon de répondre à ces trois objectifs, il est essentiel de tirer des leçons du passé qui a vu des évaluations nationales prétendre remplir conjointement plusieurs fonctions. Il apparaît important de réaffirmer, à la suite du Haut Conseil de l'éducation [2011] dans son bilan des résultats de l'école, que *« il n'est pas de bonne méthode de confondre deux types d'évaluations : d'une part les évaluations dans la classe dont l'enseignant a régulièrement besoin pour adapter son enseignement en fonction des acquis de ses élèves, d'autre part une évaluation nationale destinée au pilotage du système éducatif »*.
Le premier objectif devrait pouvoir être réalisé au travers d'outils pédagogiques proposés au niveau national ou académique, du type des anciennes évaluations CE2 et sixième, possiblement au début des nouveaux cycles, soit début de CM1 et début de cinquième, si l'on souhaite des évaluations exhaustives, grâce à des banques d'outils si l'on veut que les enseignants disposent d'outils utilisables toute l'année.

Ce type d'évaluations serait légitimement à la charge des instances pédagogiques du Ministère ou des académies.
Pour remplir le deuxième objectif, le Ministère dispose des divers outils évoqués plus haut dans cet article (dispositif Cedre, LOLF, panels, enquêtes internationales). Un effort de rationalisation de la complémentarité de ces outils a été récemment entrepris. Le dispositif Cedre a vu son cycle se réduire à cinq ans et les évaluations pour les indicateurs de la LOLF sont désormais organisées selon une périodicité de trois ans (chaque année un palier est évalué en commençant par le CE1 en 2014) et ne portent plus que sur les compétences 1 et 3 du socle commun. En outre, l'engagement renouvelé de la DEPP dans sa participation aux évaluations internationales (notamment avec la reprise de l'évaluation TIMSS en 2015) témoigne de la conscience d'une nécessaire complémentarité des points de vue sur l'état et l'évolution du système éducatif. Ainsi, PISA [OCDE, 2013] a non seulement confirmé les inégalités socio-scolaires et leur accroissement établis dans Cedre mais a en outre révélé leur aggravation comparativement aux autres pays de l'OCDE.

Le troisième objectif est sans doute le plus ardu à atteindre si l'on souhaite éviter de mélanger les deux premiers objectifs en dotant les acteurs locaux d'outils de pilotage. La DEPP a engagé un partenariat avec quelques académies pour contribuer à l'élaboration d'outils d'évaluation des acquis des élèves ou des performances des établissements. Il s'agit de fournir aux cadres territoriaux des éléments leur permettant d'envisager des mesures de régulation. Les évaluations LOLF feront ainsi l'objet tous les trois ans (en début de sixième) d'échantillons académiques représentatifs permettant d'apprécier l'évolution des performances de chaque académie. Une réflexion est actuellement menée entre la DEPP et quelques responsables académiques pour trouver une solution qui permette l'exhaustivité de ces évaluations dans leur académie en surmontant d'importants problèmes conceptuels, techniques, voire politiques.

Par ailleurs, la DEPP expérimente depuis quelques années la possibilité d'une administration informatisée des évaluations, en parallèle avec des études sur la comparabilité des supports utilisés : version papier-crayon *vs* version sur écran[21]. Le développement de ces évaluations progresse mais de redoutables défis restent encore à relever, notamment ceux de la qualité des infrastructures informatiques, en particulier dans les écoles, de l'accès des établissements au haut débit, de la capacité des acteurs locaux à assurer la maintenance et le renouvellement de leur parc informatique, de l'accompagnement des établissements scolaires dans la mise en œuvre de ces évaluations, etc. Cette nouvelle génération d'évaluations devrait permettre de faciliter la mise en œuvre pratique d'évaluations standardisées, qu'elles soient à visée bilan ou diagnostique, sur échantillons ou exhaustives. En outre, la nature même des compétences qui pourront être évaluées au moyen du support numérique devra être envisagée de façon renouvelée et innovante.

21. Ces études sont indispensables pour assurer la comparabilité temporelle d'évaluations passées sur des supports de différentes natures. Ce type d'études est également à l'ordre du jour des enquêtes internationales. Ainsi, pour PISA 2015, l'enquête se déroulera entièrement sur ordinateurs dans la majorité des 70 pays participants.

BIBLIOGRAPHIE

Andreu S., Étève Y., Garcia E., 2014, « Cedre 2013 – Grande stabilité des acquis en sciences en fin d'école depuis 2007 », *Note d'information,* n° 27, MENESR-DEPP.

Baudelot C., Establet R., 2009, *L'élitisme républicain – L'école française à l'épreuve des comparaisons internationales*, Paris, Le Seuil.

Baudelot C., Establet R., 1989, *Le niveau monte – Réfutation d'une vieille idée concernant la prétendue décadence de nos écoles*, Paris, Le Seuil.

Baudelot C., Establet R., 1988, « Le niveau intellectuel des jeunes conscrits ne cesse de s'élever », *Économie et Statistique*, n° 207, Insee, p. 31-39.

Bessonneau P., Beuzon S., Boucé S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin de collège de 2004 à 2010 », *Note d'information*, n° 12.05, MENJVA-DEPP.

Bessonneau P., Beuzon S., Daussin J.-M., Garcia E., Lévy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin d'école de 2004 à 2010 », *Note d'information*, n° 12.04, MENJVA-DEPP.

Blum A., Guérin-Pace F., 2000, *Des lettres et des chiffres – Des tests d'intelligence à l'évaluation du « savoir lire », un siècle de polémiques*, Paris, Fayard.

Bonora D., Vrignaud P., 1997, *Analyse interne utilisant les modèles MRI (ou IRT) des épreuves de mathématiques dans les dispositifs troisième de 1984, 1990 et 1995 : difficulté des items et évolution des compétences*, Rapport de convention MENESR-DEP, CNAM-INETOP.

Bonora D., Vrignaud P., 1996, *Étude de l'évolution des connaissances des élèves en début de sixième, perspective psychométrique classique et perspective* MRI (ou IRT), Rapport de convention, MENESR-DEP, CNAM-INETOP.

Bottani N., Vrignaud P., 2005, *La France et les évaluations internationales*, Les rapports établis à la demande du Haut Conseil de l'évaluation de l'école, Rapport n° 16. www.ladocumentationfrancaise.fr/rapports-publics/054000359/

Bourny G., Bessonneau P., Daussin J.-M., Keskpaik S., 2010, « L'évolution des compétences générales des élèves en fin de collège de 2003 à 2009 », *Note d'information*, n° 10.22, MEN-DEPP.

Bret A., Garcia E., Roussel L., 2014, « Cedre 2013 – Sciences en fin de collège : stabilité des acquis depuis six ans », *Note d'information*, n° 28, MENESR-DEPP.

Breton X., Marc A., 2009, *Les évaluations dans l'enseignement primaire au service de la réussite scolaire – Les propositions du Groupe UMP*, Assemblée nationale, p. 8.

Brézillon G., Chollet-Remvikos P., Rebmeister B., Zelty C., 2003, « Évaluations CE2 - sixième - cinquième – Repères nationaux septembre 2002 », *Les Dossiers Enseignement scolaire*, n° 141, MJENR-DEP.

Colmant M., Daussin J.-M., Bessonneau P., 2011, « Compréhension de l'écrit en fin d'école – Évolution de 2003 à 2009 », *Note d'information*, n° 11.16, MEN-DEPP.

Colmant M., Le Cam M., 2012, « Pirls 2011 – Étude internationale sur la lecture des élèves au CM1 – Évolution des performances à dix ans », *Note d'information*, n° 12.21, MEN-DEPP.

Daussin J.-M., Keskpaik S., Rocher T., 2011, « L'évolution du nombre d'élèves en difficulté face à l'écrit depuis une dizaine d'années », *France, portrait social*, Insee, p. 137-152.

Daussin J.-M., Rocher T., Trosseille B., 2010, « L'attestation de la maîtrise du socle commun est-elle soluble dans le jugement des enseignants ? » *Éducation & formations*, n° 79, MENJVA-DEPP, p. 45-58.

Dejonghe V., Levasseur J., Alinaudm B., Peretti C., Petrone J.-C., Pons C., Thélot C., 1996, « Connaissances en français et en mathématiques des élèves des années 20 et d'aujourd'hui », *Les dossiers d'Éducation et Formations*, n° 62, MENESR-DEP.

Dessus N., Jouvanceau P., Murat F., 1996, « Les connaissances des élèves en fin de troisième générale – évolution 1984-1990-1995 », *Note d'information*, n° 96.36, MENESR-DEP.

Dickes P., Tournois J., Flieller A., Kop J.-L., 1994, *La psychométrie – Théorie et pratique de la mesure en psychologie*, Paris, PUF.

Éducation & formations, 1990, « *Évaluation CE2-sixième – Résultats nationaux septembre 1989* », n° hors-série, Paris, MENJS-DEP, 57 p.

Flieller A., 2001, « Problèmes et stratégies dans l'explication de l'effet Flynn », *in* Huteau M., *Les figures de l'intelligence* , Paris, Éditions et applications psychologiques, p. 43-66.

Flieller A., Manciaux M., Kop J.-L., 1995, *Comparaison des compétences cognitives de deux cohortes d'écoliers de 7 ans observées à vingt ans d'intervalle* (1973-1992), Rapport final, ADEPS-Nancy 2, École de Santé publique Henri-Poincaré.

Flieller A., Santigny N., Schaeffer R., 1986, « L'évolution du niveau intellectuel des enfants de 8 ans sur une période de 40 ans (1944-1984) », *L'Orientation scolaire et professionnelle*, n° 15, p. 61-83.

Garcia É., Krop J., 2013, « Cedre 2012 histoire-géographie et éducation civique : baisse des acquis des élèves de fin de collège depuis six ans », *Note d'information*, n° 13.11, MEN-DEPP.

Garcia É., Pastor J.-M., 2013, « Cedre 2012 histoire-géographie et éducation civique en fin d'école primaire : grande stabilité des acquis depuis six ans », *Note d'information*, n° 13.10, MEN-DEPP.

Haut Conseil de l'Éducation, 2011, *Les indicateurs relatifs aux acquis des élèves – Bilan des résultats de l'école-2011*.
http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/114000565/0000.pdf

Haut Conseil de l'évaluation de l'école, 2001, « Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans », avis n°2, MEN-DEP.

Haut Conseil de l'évaluation de l'école, 2003, « Éléments de diagnostic sur le système scolaire français », Avis n° 9, MEN-DEP.

IGEN 2007, *Note sur le suivi de la mise en œuvre de l'évaluation des élèves à l'entrée de la première année du cours élémentaire* (CE1), MENESR, rapport n° 2007-030, 39 p.

IGEN-IGAENR 2009, *Troisième note de synthèse sur la mise en œuvre de la réforme de l'enseignement primaire*, Paris, MENESR, Note 2009-072, juillet 2009, 28 p.

IGEN-IGAENR, 2005, *Les acquis des élèves, pierre de touche de la valeur de l'école ?* Paris, MENESR, rapport n° 2005-079, 83 p.

Jonas N., 2012, « Pour les générations les plus récentes, les difficultés des adultes diminuent à l'écrit, mais augmentent en calcul », *Insee Première*, n° 1426, Insee.

Jones L. V., Olkin I., 2004, *The nation's report card: Evolution and perspectives*, Bloomington, Phi Delta Kappa Educational Foundation.

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (2e édition), Bruxelles, De Boeck.

Le Cam M., Rocher T., Verlet I., 2013, « Forte augmentation du niveau des acquis des élèves à l'entrée au CP entre 1997 et 2011 », *Note d'information*, n° 13.19, MEN-DEPP.

Le Guen M., 1991, « L'évaluation des acquis des élèves : caractéristiques et évolution du dispositif national », *L'orientation scolaire et professionnelle*, vol. 20, n° 1, p. 39-69.

Levasseur J., 1996, « L'évaluation nationale des acquis des élèves », *Revue internationale d'éducation de Sèvres*, n° 11, CIEP, p. 101-114.

Mons N., 2008, « Évaluation des politiques éducatives et comparaisons internationales », *Revue française de pédagogie*, n° 164, ENS Éditions, p. 5-13.

Mons N., 2007, *Les nouvelles politiques éducatives – La France fait-elle les bons choix ?* Paris, PUF.

Mullis I., Martin M., Foy P., Arora A., 2012, *TIMSS 2011 international results in reading*, Chestnut Hill, MA, TIMSS & PIRLS International Study Center, Boston College.

National Commission on Excellence in Education, 1983, *A Nation at risk: the imperative for educational reform*, Washington D.C.

OCDE, 2013, *Résultats du PISA 2012 : savoirs et savoir-faire des élèves*, vol. 1 à 5, Paris.

Pair C., 2001, *Forces et faiblesses de l'évaluation du système éducatif en France*, Rapport n° 3, rapport établi à la demande du Haut Conseil de l'évaluation de l'école. http://www.ladocumentationfrancaise.fr/rapports-publics/024000206/

Peretti C., Petrone J.-C., Thélot C., 1996, « L'évolution des compétences scolaires des meilleurs élèves depuis 40 ans », *Les dossiers d'Éducation & formations*, n° 69, MENESR-DEP.

Piéron H., 1963, *Examens et docimologi*e, Paris, PUF.

Pons C., 1996, « Connaissances en français et en calcul des élèves des années 20 et d'aujourd'hui », *Note d'information*, n° 96.19, MENESR-DEP.

Rocher T., 2012, « Comment assurer la comparabilité des scores issus d'évaluations nationales annuelles et exhaustives ? Comparaison de différentes méthodes d'ajustement des métriques (*equating*) », *Actes du 24e colloque international de l'ADMÉE-Europe*, Luxembourg, papier présenté au 24e colloque international de l'ADMÉE-Europe (admee2012.uni.lu).

Rocher T., 2010, « La performance de l'école primaire : quelques résultats récents tirés de l'évaluation des acquis des élèves », *Administration et Éducation*, n° 125, AFAE, p. 43-50.

Rocher, T., 2008a, « Lire, écrire, compter : les performances des élèves de CM2 à vingt ans d'intervalle (1987-2007) », *Note d'information*, n° 08.38, MEN-DEPP.

Rocher T., 2008b, « Que nous apprennent les évaluations internationales sur le fonctionnement des systèmes éducatifs ? Une illustration avec la question du redoublement », *Éducation & formations*, n° 78, p. 63-68, MEN-DEPP.

Rocher T. 2008c, « La détermination de standards minimaux dans le cadre d'indicateurs de résultats : méthodologie, intérêt, utilité », *Mesure et évaluation en éducation*, vol. 31, n° 2, ADMÉE Canada, p. 75-91.

Rocher T., Le Donné N., 2012a, « Les aspirations professionnelles des élèves de 15 ans dans 57 pays : ambition et réalisme », *L'orientation scolaire et professionnelle*, vol. 41, n° 3, p. 439-468.

Rocher T., Le Donné N. 2012b, « Les difficultés de lecture en début de sixième – Évolution à dix ans d'intervalle (1997-2007) », *Éducation & formations*, n° 82, MEN-DEPP, p. 31-37.

Salines M., Vrignaud P., 2001, *Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans*, Rapport n° 2, rapport établi à la demande du Haut Conseil de l'évaluation de l'école.
http://www.ladocumentationfrancaise.fr/rapports-publics/024000205/index.shtml

Thélot C., 1992, « Que sait-on des connaissances des élèves ? », *Les dossiers d'Éducation & formations*, n° 17, MEN-DEP.

Vourc'h R., Rivière J.-P., De La Haye, F., Gombert J.-E., 2014, « Journée défense et citoyenneté 2013 : des difficultés en lecture pour un jeune français sur dix », *Note d'information*, n° 12, MENESR-DEPP.

Zwick R., 1992, "Statistical and psychometric issues in the measurement of educational achievement trends: examples from the National Assessment of Educational Progress", *Journal of Educational Statistics*, vol. 17, No. 2, AREA, p. 205-218.