Apprécier et certifier les acquis des élèves en fin de collège : brevet et évaluations-bilans

Michel SALINES
Inspecteur d'académie honoraire

Pierre VRIGNAUD
Chercheur
Institut national d'étude du travail
et d'orientation professionnelle,
Conservatoire national
des arts et métiers

Juin 2001

Rapport établi à la demande du
Haut Conseil de l'évaluation de l'école

Le Haut Conseil de l'Évaluation de l'École a souhaité examiner les moyens dont on dispose pour évaluer les acquis des élèves à la fin de la classe de 3ème, sous la forme de l'examen du Brevet des Collèges et des évaluations-bilan.

La dernière année de collège représente, dans les faits, la fin de la scolarité obligatoire. Deux activités d'évaluation sont conduites en fin de collège. L'une concerne directement chaque élève : l'examen conduisant au diplôme national du brevet ; l'autre vise les élèves dans leur ensemble : l'évaluation-bilan. Les finalités de ces deux entreprises sont différentes. Le brevet certifie les acquis ou les compétences de chaque élève, à la fin d'un parcours de neuf années de scolarité obligatoire dans un cursus unique. Les évaluations-bilans en fin de troisième fournissent une image des acquis et compétences de l'ensemble des élèves à cette dernière étape de la scolarité commune. Elles apportent des éléments de base à un observatoire qui permet d'analyser le système éducatif, pour le piloter et pour alimenter le débat public sur l'école.

Le lecteur trouvera, dans le rapport qui suit, les résultats du travail d'enquête et d'analyse que nous avons mené sur ces deux formes de dispositifs d'évaluation.

Remerciements :

Nous remercions toutes les personnes du Haut Conseil de l'Évaluation de l'École, de l'Inspection Générale de l'Education Nationale, de la Direction de la Programmation et du Développement, les responsables des services d'éducation des départements de Corrèze et de Seine et Marne, ainsi que les collègues des différents organismes étrangers d'évaluation des acquis des élèves qui ont été d'une grande disponibilité pour répondre à nos questions et nous rendre accessibles publications, documents et données.

TABLE DES MATIERES

I - LE DIPLOME NATIONAL DU BREVET

Rapport conduit par Michel SALINES

1 - HISTORIQUE

Le diplôme national du Brevet a été créé par le décret du 23 janvier 1987. Les textes pris en application de ce décret ont été modifiés en 1999 par un arrêté du 18 août et une note de service du 6 septembre.

Ce diplôme, ouvert aux élèves des classes de 3ème des collèges, c'est à dire en fin de scolarité obligatoire, est l'héritier direct du Brevet élémentaire créé sous la 3ème République, et du Brevet d'études du Premier cycle (BEPC) créé en 1948.

1.1 - RAPPEL

Le Brevet élémentaire, qui a existé jusqu'à la fin des années 50, sanctionnait les études en fin de 3ème. Dès sa création, il est considéré comme un diplôme de capacité pour l'enseignement du premier degré. Il conserve ce caractère, même après la mise en place du Brevet Supérieur, titre en principe requis, pour se présenter aux épreuves du CAP d'instituteur. Le Brevet est apparu très rapidement comme le diplôme de fin d'études, réservé aux élèves ne fréquentant pas le lycée (il n'était pas du tout préparé dans les lycées et les élèves ne s'y présentaient pas), mais les collèges modernes et surtout les Cours Complémentaires, dont il faut rappeler qu'ils furent le premier pas vers un enseignement « populaire » de second degré et donc la première institution de démocratisation de l'enseignement secondaire.

Le Brevet était un examen difficile, organisé au plan départemental. Sa valeur était reconnue socialement et dans le monde du travail, car il garantissait un niveau de base satisfaisant dans les différents savoirs constitutifs d'un socle culturel commun (on parlerait aujourd'hui de culture commune). Il portait sur toutes les disciplines à l'exclusion des langues vivantes, non enseignées dans les Cours Complémentaires, où il était, pour l'essentiel préparé.

Le Brevet était exigé pour le recrutement des cadres intermédiaires de la fonction publique : Postes, Finances, Police et autres administrations. Les employés subalternes étaient eux, recrutés au niveau du certificat d'études primaires.

Après la guerre de 39-45, dans le droit fil des réflexions du Plan Langevin-Wallon, fut créé le BEPC. Le Brevet continua à exister pendant encore quelques années et, dans les Cours Complémentaires, il était fréquent de présenter les meilleurs élèves au Brevet, en même temps qu'au BEPC. Il a subsisté jusqu'en 1959. Le BEPC était un examen plus souple, comportant de nombreuses options, passé aussi bien par les élèves des Cours Complémentaires que par ceux des lycées. Il sanctionnait le niveau atteint par les élèves en fin de classe de 3ème, mais il ne bénéficia jamais de la reconnaissance sociale et professionnelle attachée au Brevet élémentaire. Le BEPC, examen national comme le Brevet, comportait des épreuves écrites et des épreuves

orales auxquelles il fallait être admissible : c'était en quelque sorte le Baccalauréat du premier cycle du second degré. Tous les élèves ne l'obtenaient pas et le pourcentage de reçus pouvait être faible (parfois moins de 60% dans certains départements, voire beaucoup moins dans certains établissements).

Le BEPC fut modifié à plusieurs reprises, mais il avait perdu une grande partie de sa signification et de son prestige, car il était devenu à la longue une sorte d'attestation de fin d'études, accordée à tous les élèves de 3ème. Pour tenter de le moderniser en restaurant son image, il fut remplacé en 1987 par le Diplôme national du Brevet.

De ce bref historique nous retiendrons plusieurs caractéristiques :

- Le Brevet n'a jamais été utilisé comme passeport pour la poursuite d'études dans le second cycle des lycées.
- S'il est, depuis le début, un véritable diplôme de fin d'études obligatoires il faut reconnaître que progressivement il a perdu ce caractère dans la mesure où la plupart des jeunes poursuivent aujourd'hui leurs études au-delà de la classe de 3ème. (Soulignons quand même que plus de 40 % des élèves de 3ème ne s'orientent pas vers des études générales. Un certain nombre de « savoirs de base » ne leur seront plus proposés ultérieurement. Pour eux donc, la classe de 3ème, joue le rôle d'une véritable classe de fin d'études).
- Sa reconnaissance sociale n'a été forte et sa valeur n'a été reconnue que jusqu'aux années 50.
- Son organisation a toujours été départementale.
- Les résultats des élèves n'ont jamais ou rarement été utilisés – jusqu'à ces dernières années – comme indicateurs collectifs de pilotage du système.

Ainsi, le Brevet a perdu, sa reconnaissance sociale, sa valeur individuelle (mesure du niveau de l'élève et prédiction de la réussite ultérieure), sans acquérir de valeur collective (indicateur de « valeur ajoutée » des établissements).

1.2 - LES TEXTES

Le décret de 1987 qui crée le Diplôme national du Brevet amende le précédent texte de 1976 qui portait aménagements du BEPC.

Le décret de 1987 est innovant puisqu'il instaure 3 séries : collège, technologique (élèves des classes de 3ème technologique) et professionnelle (élèves des lycées professionnels et agricoles).

Les textes de 1987 ont été modifiés en 1999 comme on l'a vu plus haut. Ces textes récents témoignent de la volonté ministérielle de « recadrer » les pratiques face à un certain nombre de dérives observées dans l'attribution de ce diplôme, d'une part, et de réaffirmer son caractère national, d'autre part.

L'arrêté de 1999 :

- précise les coefficients des épreuves écrites terminales et des notes du contrôle continu. Il faut relever que toutes les disciplines sont prises en compte dans le contrôle continu (y compris l'EPS et les enseignements artistiques) ;
- renforce le caractère académique du choix des sujets et des barèmes de correction, en instituant une commission académique. (Dans la pratique, on a regroupé depuis l'année dernière les académies – comme pour le baccalauréat – en quatre groupes d'académies, et les sujets deviennent progressivement inter académiques).

La note de service pour sa part :

- détaille les conditions de choix des sujets, de leur essai préalable et du contrôle de leur qualité, de la correction des copies, de l'harmonisation des notes ;
- précise les modalités de prise en compte des résultats acquis en cours de scolarité en classe de 4ème et de 3ème, et le mode d'élaboration des notes de contrôle continu, pour lesquelles seront seules comptabilisées celles des contrôles ponctuels et des résultats à des épreuves communes inter-classes. Elle en exclut les notes acquises au cours d'exercices d'entraînement ou d'acquisition et insiste sur la nécessaire harmonisation de ces notes dans chaque discipline, au sein de chaque établissement ;
- attire par ailleurs, l'attention des enseignants sur la nécessité de prendre en considération dans chaque discipline les capacités d'expression orale des élèves ;
- définit avec beaucoup de précision la nature et le contenu des trois épreuves de contrôle terminal : Français, Mathématiques, Histoire et Géographie.

Ces textes qui visent à garantir le niveau d'exigences du diplôme et sa cohérence nationale, ne prévoient, par contre, aucune stratégie de mise en oeuvre ou de suivi de leur application dans les académies, aucune remontée nationale des sujets des épreuves et des résultats chiffrés. De même, ils ne prévoient aucune action académique de formation des professeurs à la pratique délicate du contrôle continu.

On peut le regretter.

2 - LES ETUDES DISPONIBLES AU MINISTERE DE L'EDUCATION NATIONALE

Nous n'avons trouvé que peu d'études disponibles au plan national : tout se passe comme si le Brevet n'intéressait pas le Ministère. On a relevé ce phénomène avec quelque étonnement car il s'agit bien – répétons-le – d'un diplôme national, géré par des textes fréquemment réactualisés. Le rédacteur de l'étude « **Le collège, 7 ans d'observation et d'analyse** », (Hachette, 1999), citant l'Inspection générale, souligne que les résultats du Brevet sont « difficiles à interpréter, ne serait-ce que par suite des grandes difficultés dans les modes d'appréciation du contrôle continu. Ils doivent être accueillis avec prudence… Ils méritent donc d'être relativisés et ne devraient pas mobiliser à l'excès l'attention des établissements et des enseignants. Bien des collèges, pourtant, tombent dans ce travers et dans certains d'entre eux, où la liaison collège-lycée reste à développer, on se préoccupe plus des résultats au brevet que de la préparation des élèves à leurs études de second cycle ».

Faut-il en conclure que pour certains responsables de l'Education nationale, ce diplôme n'a plus d'importance ?

2.1 - LES TRAVAUX DES INSPECTIONS GÉNÉRALES

Nous n'avons trouvé qu'une seule étude, mais très intéressante, de 1991 dans le rapport : **« L'évaluation des acquis des élèves à la fin des cycles des apprentissages »** (voir annexe 5).

Cette étude a été réalisée sur un panel de 80 collèges dans 3 académies par l'Inspection générale de l'Education nationale. Un chapitre important y est consacré à la notation au Diplôme national du Brevet. Les constats essentiels sont les suivants :

- « le Diplôme national du Brevet n'est plus désormais conçu comme une sanction de la scolarité obligatoire mais comme une étape dans un cursus rendant compte du degré d'acquisition des connaissances, l'obtention des connaissances attestant le franchissement d'un seuil » ;
- « des études approfondies ont permis d'observer une inégalité des taux de réussite entre les départements et les établissements, ce qui ne peut s'expliquer par les seules différences de niveau des élèves. On n'est donc pas encore parvenu à une rigueur suffisante dans l'application de critères de notation définis de façon précise en fonction d'objectifs à atteindre et donc à une régulation satisfaisante du système de certification des acquis. »
- « ces constats conduisent à s'interroger sur la rigueur, la clarté et l'équité des résultats du brevet ».

Par ailleurs les Inspecteurs généraux constatent – et ce n'est pas nouveau – que les résultats du Brevet ne sont pas pris en compte pour l'orientation des élèves en fin de 3ème. Les notes utilisées par les conseils de classe pour déterminer le passage des élèves en seconde, sont très exactement les notes du contrôle continu du Brevet. Or, notent-ils, « ce ne sont pas les résultats du contrôle continu qui permettent de formuler un bon pronostic, mais les épreuves écrites qui ne sont jamais utilisées puisqu'elles sont connues après les décisions d'orientation ».

A la fin du rapport, les Inspecteurs généraux demandent que la rigueur du contrôle continu soit améliorée, en fixant au niveau national, pour « chaque discipline, des références précises sur ce que l'on doit exiger des élèves ». Ils proposent aussi de réduire et de réguler les conditions et les modalités du repêchage.

Ils constatent que le diplôme ne retrouvera « sa pleine efficacité que dans la mesure où conformément aux objectifs initiaux il constitue un véritable instrument d'évaluation qui permette aux collèges d'apprécier les résultats obtenus en les comparant à ceux des autres établissements, et aux professeurs de seconde de prévoir des remises à niveau en fonction des lacunes identifiées ». Pour la première fois – selon nous – apparaît l'idée d'utiliser les résultats du Brevet comme instruments de pilotage des établissements et du système.

Dans son rapport de 1993, l'Inspection générale de l'Administration de l'Education nationale, pour sa part, souligne très brièvement, qu'il n'y a pas lieu de remettre en cause l'existence du Brevet dont l'intérêt est à la fois pédagogique et psychologique.

2.2 - A LA DIRECTION DES ENSEIGNEMENTS SCOLAIRES

Nous n'avons rien trouvé à la Direction des établissements scolaires : ni études ni analyses sur le fonctionnement de cet examen, ni même de remontée des résultats des académies ou des départements.

Ce travail n'existe que dans les Inspections académiques où les résultats sont souvent analysés et diffusés aux établissements.

On éprouve un peu le sentiment, d'un examen marginalisé, privé de véritable considération et qui survit indépendamment des décideurs, grâce au travail et à la considération des acteurs de terrain.

3 - OBSERVATION DES PRATIQUES

3.1 - COMMENT LE BREVET FONCTIONNE-T-IL REELLEMENT ?

Pour le savoir nous nous sommes rendus dans deux départements différents par leurs structures, leur géographie, leurs traditions : la Seine et Marne et la Corrèze. Bien entendu ce compte-rendu ne prétend pas être représentatif, au sens statistique du terme, mais ces études de cas peuvent éclairer le débat.

Dans chacun des deux départements nous avons rencontré l'Inspecteur d'Académie et les services des examens des Inspections académiques ainsi que les équipes de deux collèges : principal et professeurs de classe de 3ème en Lettres, Mathématiques, Histoire et Géographie. Un guide d'entretien sommaire leur avait été adressé quelques jours avant la rencontre.

Quelles conclusions après ces visites ?

- **Le Brevet est pleinement l'affaire de l'Inspecteur d'Académie qui en assume personnellement et très activement la responsabilité**, pour ce qui touche à :
    - l'organisation,
    - la composition des jurys,
    - la gestion des résultats et des décisions de repêchage,
    - la centralisation des notes, les synthèses, l'exploitation des résultats.

- **Les sujets**

  Ils sont en principe inter-académiques : dans chacun des 4 groupes d'académies, les sujets de contrôle terminal sont proposés à tour de rôle par chaque académie. Ils sont préparés par les IPR et une commission disciplinaire académique. Dans certaines matières un cadrage national a été mis en place : en Lettres et en Histoire et Géographie par exemple, la réflexion est engagée depuis deux ans, sur la conformité des sujets avec les programmes de la classe de 3ème et sur leur valeur prédictive de la réussite dans la poursuite des études au lycée.

- **Le contrôle continu**

  A priori la mise en place du contrôle continu – partie intégrante des résultats permettant d'attribuer le diplôme – est bien admise par tous les responsables et par les professeurs rencontrés qui souhaitent tous son maintien, **mais les textes de 1999 qui recadrent les pratiques et qui apportent des instructions précises pour la mise en œuvre du contrôle continu, ne sont pas connus dans les établissements et donc restent inappliqués**.

  Le Ministère, les rectorats et les IA les ont diffusés, mais comme aucune stratégie d'information, et de formation n'avait été définie pour accompagner le texte, ils n'ont donné lieu à aucune animation sur le terrain ni à aucune initiative visant à les faire connaître et appliquer. Du coup ils ne sont pas plus lus que nombre d'autres textes réglementaires. Tout se passe parfois dans notre administration comme si la seule diffusion d'un texte au bulletin officiel, suffisait à en garantir l'application. Nous savons très bien que ce n'est pas suffisant, surtout lorsqu'il s'agit de modifier des pratiques ou de mettre en œuvre des mesures pédagogiques complexes. Ce type de stratégie ne permet pas toujours au système éducatif d'évoluer en corrigeant ses erreurs ou en palliant ses difficultés. Pour mémoire, la mise en œuvre du contrôle continu dans l'enseignement technique a donné lieu à des impulsions

politiques claires, et à des années de formation des personnels : ce fut sûrement la clef du succès actuel.

Nous avons constaté – comme l'Inspection générale en 1991 – le plus grand désordre dans la mise en œuvre du contrôle continu : des différences très importantes d'établissement à établissement dans un même département, et bien souvent au sein du même établissement, entre les disciplines, voire même au sein d'une même discipline.

Or, il faut le souligner, les professeurs attachent la plus grande importance à la notation des élèves et le plus grand sérieux à la mise en œuvre du contrôle continu. A ce sujet, plusieurs études montrent que les élèves, à l'intérieur d'une division, sont classés schématiquement de la même manière par les évaluations nationales, les épreuves terminales, et le contrôle continu. Par contre, les échelles de notation ne sont pas identiques, d'un instrument à l'autre, d'un établissement à l'autre. Si tel était le cas, dans les classes de collèges rattachées à de grands lycées parisiens, la note la plus basse serait 12, et dans les classes de collège en ZEP, la note la plus haute serait très inférieure. Sommes-nous prêts à accepter ce phénomène ?

En ce qui concerne le contrôle continu, le fait que soient associées les notes obtenues en classe de 4ème et en classe de 3ème, rend la mesure plus solide. Malgré cela, nous avons constaté un très fort décalage entre les notes de contrôle continu et celles de contrôle terminal dans les matières où la comparaison est possible. Cette incohérence pose le problème de la validité des résultats à l'examen et aussi celle d'une évaluation des élèves, qui se borne souvent à la notation des manques par rapport à une norme propre à chaque enseignant, en omettant d'évaluer les acquis.

- **L'utilisation des résultats individuels et collectifs des élèves**

Les résultats du Brevet sont utilisés régulièrement par les Inspecteurs d'Académie rencontrés qui les regroupent dans une analyse départementale, collège par collège, et qui les intègrent dans d'excellents tableaux de bord par établissement, utilisés comme outil de pilotage des projets pédagogiques de chacun des collèges. Les principaux les communiquent souvent aux équipes de professeurs comme support de la réflexion collective.

En l'absence de modèles d'analyse qui pourraient être élaborés au plan national, (par la DPD par exemple), chaque département a créé ses propres modes de traitement des informations et on constate une assez forte hétérogénéité d'un département à l'autre. Relevons aussi que les services statistiques des rectorats, dont on connaît la compétence, ne sont malheureusement pas mobilisés pour cette tâche.

Les principaux utilisent ces résultats, mais parfois de façon assez sommaire et ponctuelle : il y a peu de réflexions collectives sur la mise en œuvre du contrôle continu – on l'a vu. Ce travail collectif intervient en général en amont de l'examen, dans la mise en place d'épreuves communes ou de brevets blancs dont l'usage se développe de plus en plus.

Enfin, comme le souligne le rapport de l'Inspection générale en 1991, les notes des contrôles terminaux et les résultats des élèves au Brevet ne sont jamais utilisés dans les procédures d'orientation. On note donc partout, un certain décalage entre les décisions d'orientation en seconde Générale et Technologique et les résultats au Brevet : nombre d'élèves qui ne sont pas reçus au Brevet sont pourtant orientés en classe de seconde…

- **Une image forte**

  Unanimement, tous nos interlocuteurs ont souhaité le maintien du Brevet en relevant qu'il garde une image forte chez les parents, les élèves et les enseignants, comme si perdurait dans l'inconscient collectif, le prestige du Brevet élémentaire de nos grands-parents… Pour autant, ils n'écartent pas l'idée de certains aménagements. Tous regrettent, par exemple, que la Langue Vivante 1 ne fasse pas partie des épreuves terminales ; ils souhaitent que soient revus les coefficients de certaines disciplines du contrôle continu.

  Les raisons qui militent en faveur du maintien du Brevet et qui sont évoquées, sont les suivantes :
    - c'est un rite initiatique qui apprend aux élèves à gérer le stress,
    - il responsabilise les élèves et leur donne de l'importance : leur dignité est reconnue,
    - il légitime le travail des professeurs,
    - il permet un bilan de l'élève,
    - c'est le nécessaire diplôme de fin d'études obligatoires,
    - les résultats obtenus jouent un rôle important pour l'image du collège dans son environnement.

3.2 - LE BREVET : UN EXAMEN POLYMORPHE

Le Brevet comporte, on l'a vu, 3 séries :
- la série générale,
- la série technologique,
- la série professionnelle.

La série générale concerne tous les élèves en fin de 3ème : les sujets du contrôle terminal sont communs.

La série technologique concernait jusqu'à ces dernières années, les élèves de 3ème technologique de collège et de lycée professionnel. Les sujets ne sont pas identiques à ceux de la série générale même si les épreuves sont sensiblement les mêmes (la note de technologie de contrôle continu est dans cette série affectée du coefficient 2).

La série professionnelle ne concerne, elle, que les élèves scolarisés en lycée professionnel dans les classes de préparation au CAP et BEP, et sa mise en œuvre est diverse selon les départements. Le diplôme fait d'ailleurs double emploi, avec le CAP ou le BEP :
- les élèves issus de classes de 4ème dites faibles, scolarisés - en principe - en section CAP, passent un Brevet « adapté » en 6 épreuves.
- les autres élèves de 3ème professionnelle passent le Brevet, option professionnelle.

En lycée professionnel se présentent à l'examen, les élèves qui ne sont pas, à l'entrée, titulaires du Brevet. Les chefs d'établissement, consultés, nous ont indiqué que le bénéfice psychologique est très important, pour ces élèves qui ont eu une scolarité difficile en collège. C'est pour eux – enfin – un succès qui les motive pour la suite de leurs études. Il faut ajouter que certains concours du tertiaire exigent le Brevet ou le BEP, (par exemple les concours de recrutement d'aide-soignante ou d'auxiliaire de puériculture).

On est tenté de se demander si cette pratique ne constitue pas un leurre ou une illusion, en attribuant un diplôme d'enseignement général à des jeunes en formation professionnelle, formation sanctionnée par le CAP et le BEP. On peut difficilement laisser croire que ces différents examens sont équivalents. On pourrait imaginer plutôt que les élèves ayant échoué en classe de 3ème repassent – s'ils le souhaitent – l'examen en candidats libres, l'année suivante, rien n'interdisant d'ailleurs aux établissements d'accueil de les y préparer par des compléments facultatifs.

3.3 - LES SUJETS

Il semblait nécessaire de mener une analyse aussi précise que possible des sujets des épreuves terminales sur deux plans :

- les sujets sont-ils bien en phase avec les programmes du collège en général, et ceux de la classe de 3ème en particulier ?
- les résultats obtenus par les élèves et leurs performances à cet examen, constituent-ils une bonne prédiction de leur succès ultérieur ?

Cette analyse, rapide, compte tenu du temps disponible, a été menée par l'Inspection générale qui a accepté de nous confier ses conclusions pour les trois épreuves terminales : Français, (Mme Weinland), Mathématiques (M. Fort), Histoire et Géographie (M. Hagnerelle). On les trouvera respectivement en annexes 1, 2 et 3.

Tous s'accordent à reconnaître que, pour diverses raisons, avant les textes de 99, les sujets étaient inadaptés aux objectifs et aux contenus de l'enseignement au collège tels qu'ils ont été définis pour chaque discipline, et étaient souvent obsolètes : morcellement inacceptable en Français où tout a été mis en œuvre ces dernières années pour décloisonner les différentes activités ; inadaptation aux contenus de l'enseignement en Histoire et Géographie (on attendait des candidats des mini-dissertations avec des exigences énormes sans rapport avec les objectifs et les contenus qui sont désormais conseillés ; absence d'épreuve d'Instruction civique. En Mathématiques, on note que malgré les instructions de 99, certaines parties du programme de 3ème ne sont jamais évaluées (différentes selon les académies).

Tous relèvent un suivi inégal des procédures de correction et d'harmonisation et l'absence d'un vrai pilotage pédagogique dans la mise en œuvre de cet examen.

Un effort important a été engagé depuis la parution des textes de 99 : comme l'un des inspecteurs généraux le note sans ambages : « à l'occasion de ces textes et des nouveaux programmes, l'Inspection générale se ré-intéresse au Brevet ». En Français et en Histoire et Géographie des initiatives nationales ont été prises dès l'année scolaire 1999-2000 pour corriger les carences observées dans les sujets précédents et utiliser les résultats comme révélateurs des difficultés des élèves. Ainsi, en Histoire et Géographie, depuis la session dernière, les nouvelles épreuves qui correspondent mieux aux nouveaux apprentissages, révèlent des manques de compétence des élèves sur certains points importants des programmes : la comparaison entre deux documents par exemple. En Instruction civique, où le niveau est faible, les épreuves font apparaître la difficulté à bâtir un raisonnement logique (à cause d'un déficit de vocabulaire). On le voit, nous assistons à une véritable modification de l'attitude au plan national, non seulement dans le choix des sujets, mais dans l'utilisation systématique et raisonnée des résultats du contrôle terminal. En mathématiques, des analyses sont en cours dans plusieurs académies, qui mettent en évidence les carences des sujets et des candidats.

Les Inspecteurs généraux recommandent que soit assuré un véritable pilotage pédagogique de l'examen : national dans sa conception, en développant et en institutionnalisant les groupes de travail qui existent déjà ; académique dans sa réalisation, en impliquant mieux les corps d'inspection et la commission académique prévue par les derniers textes. Il serait sans doute bon qu'une évaluation des pratiques dans chacune des académies soit menée rapidement.

4 - LES CONSTATS

On peut donc, sans conteste, affirmer que le Diplôme national du Brevet remplit aujourd'hui de façon souvent imparfaite, les fonctions qui lui ont été assignées par la tradition scolaire et les textes qui le régissent.

4.1 - UN DIPLOME NATIONAL DE FIN D'ETUDES OBLIGATOIRES ?

Nous nous sommes interrogés sur sa validité en tant que diplôme de fin d'études obligatoires, car il nous semble que l'acquisition des connaissances de base légitimement exigibles d'un citoyen à la fin de la scolarité obligatoire n'est que maladroitement mesurée par cet examen (toutes choses égales, le certificat d'études primaires de la 3ème République – sans l'idéaliser – constituait, à certains égards, une meilleure garantie), et cela pour plusieurs raisons :

- Le rôle du collège (dernier cycle d'une scolarité obligatoire, commune à tous les enfants ? premier cycle du lycée ?) et les objectifs de fin de scolarité obligatoire ne sont pas définis de façon précise : s'il s'agit – comme on le dit depuis quelques années – de définir les savoirs de base de « l'honnête homme de ce début de siècle », où sont-ils définis et qui les définit ? De fait, tout se passe comme si on s'en remettait aux établissements, aux professeurs, pour le faire, seuls, chacun à sa place. Dans ces conditions, il ne faut pas s'étonner que chacun agisse en fonction de ses propres exigences et parfois de façon strictement disciplinaire. Nous avons rappelé plus haut que l'on note souvent les élèves par rapport à leurs manques (référés à des normes individuelles variables, chacun ayant sa propre idée du niveau exigible). On ne peut donc pas parler d'évaluation, aucune comparaison fiable n'est possible entre élèves ou entre établissements, et on aboutit nécessairement à des iniquités.

- Le niveau de l'élève par rapport à des objectifs comportementaux (souvent à la base de la réussite ultérieure, aussi bien dans la poursuite d'études que dans la vie sociale et professionnelle) n'est pas véritablement évalué (même si les textes de 99 prévoient son évaluation dans le cadre du contrôle continu). Il en est ainsi de l'aptitude à l'expression et à la communication, par exemple.

On peut s'interroger sur la valeur du Brevet actuel dans le monde du travail et de l'entreprise : « tout le monde s'en moque ! » nous a indiqué le responsable de la formation du MEDEF, car il ne permet pas d'affirmer que son titulaire dispose – selon ses propos – du «kit de base » garantissant qu'il pourra tenir sa place dans l'entreprise, être formé et progresser.

En conclusion, lorsque les réflexions engagées par le Ministre sur les objectifs du collège et le niveau exigible des élèves à l'issue de la classe de 3ème auront abouti, il sera alors possible de définir les épreuves d'un Brevet rénové, qui remplisse la fonction – nécessaire, selon nous – de diplôme de fin d'études obligatoires.

4.2 - UN OUTIL DE MESURE DU NIVEAU DE L'ELEVE DANS DIFFERENTES DISCIPLINES ?

Nous nous sommes interrogés aussi, sur la validité du Brevet, outil de mesure du niveau de l'élève dans les différentes disciplines des programmes du collège. Au cours de nos visites sur le terrain nous avons constaté que le Brevet mobilise de façon très active toutes les énergies. Il est traité très sérieusement par tous, responsables, enseignants, candidats, et cependant, il fonctionne mal.

Nous avons observé partout de fortes différences entre les notes de contrôle continu et de contrôle terminal dans les disciplines où la comparaison est possible. A cet égard, nous n'avons guère progressé depuis 1991. Si l'idée d'associer le contrôle continu des connaissances aux épreuves du contrôle terminal nous semble excellente, encore faudrait-il qu'il soit mis en œuvre de façon relativement homogène et rigoureuse au sein de chaque établissement, et d'un collège à l'autre. Pour le moment ce n'est pas le cas.

L'Inspection générale a commencé à recadrer le choix des sujets, mais nous avons relevé que l'harmonisation de la notation n'était pas encore satisfaisante et que des différences notables persistent d'un département à l'autre, d'un collège à l'autre.

Dans ces conditions certains se demandent si le Brevet est toujours un diplôme national.

Dans un autre domaine, on peut regretter que les langues vivantes (au moins une, la première langue) soient absentes des épreuves de contrôle terminal et que l'informatique n'apparaisse pratiquement nulle part dans l'examen (elle n'est évaluée que dans le cadre de l'enseignement de la technologie) alors même que l'on vient de créer le Brevet informatique et que les études récentes de l'Inspection générale sur la mise à niveau en informatique en classe de seconde révèlent une faiblesse générale des élèves dans ce domaine, avec une très grande hétérogénéité des résultats, souvent liée à l'origine sociale et culturelle des élèves, (comme si cette formation n'existait efficacement qu'en dehors des établissements scolaires) ! Il y a là un vrai problème de société : les citoyens français du XXIème siècle possèderont-ils les outils de communication et d'ouverture aussi fondamentaux pour eux comme pour le pays, que le « lire, écrire, compter » du siècle dernier ?

Les disciplines artistiques ne sont pas non plus à leur vraie place, alors que l'évaluation du niveau et des aptitudes des élèves y est possible comme le montre la pratique de certains pays européens (les Pays-Bas ont créé, par exemple, des épreuves d'évaluation en musique, qui semblent fiables et faciles à utiliser collectivement).

Enfin, même si la tradition est solidement installée (depuis toujours, nous l'avons vu) de ne pas utiliser les résultats des élèves au Brevet comme « passeport » pour la poursuite d'études, est-il normal que les notes qui leur sont attribuées, en particulier lors des épreuves du contrôle terminal (corrigées anonymement et par des correcteurs extérieurs), ne soient, ni prises en compte au moment de la préparation de l'orientation et du dialogue avec les élèves et leur famille, ni portées à la connaissance des professeurs des lycées d'accueil, alors même que la mise en œuvre de l'évaluation à l'entrée en classe de seconde est pratiquée et utilisée de façon très inégale ?

4.3 - UN OUTIL DE PILOTAGE DES ETABLISSEMENTS ET DU SYSTEME EDUCATIF ?

On peut s'étonner enfin, qu'aucun travail de remontée et d'exploitation des résultats des élèves ne soit mené, ni au plan national, ni au plan académique. Nous l'avons vu, même si des évaluations nationales périodiques sont mises en œuvre par la DPD, des informations aussi précieuses que les résultats des élèves au Brevet, pourraient être utilisées. Des expériences intéressantes sont menées en ce sens dans certains départements qui fournissent aux

établissements les éléments qui, en confrontant le niveau des élèves en 6ème à leurs résultats au Brevet, permettent de mettre en évidence « la valeur ajoutée » dont ils ont bénéficié. Au fond, le Brevet a gardé, malgré les textes de 87 et de 99, un peu de son caractère d'examen « maison » contre lequel ces textes réagissaient pour maintenir son caractère de diplôme national de fin d'études : en quelque sorte le Certificat d'études primaires du XXIème siècle.

En conclusion, le Brevet est selon nous un examen nécessaire. Il est unanimement souhaité par tous les acteurs. Il fonctionne mal aujourd'hui car il répond maladroitement aux fonctions qu'il est censé assumer. Curieusement, il est aujourd'hui d'autant plus pris en considération qu'on se rapproche du terrain : département, collège, classe, élèves et familles…
S'il garantissait la maîtrise d'une culture de base large (qui reste à définir), il pourrait être pris en considération dans le monde du travail et de l'entreprise, comme il le fut jadis.

5 - LES PROPOSITIONS

5.1 - MAINTENIR UN EXAMEN NATIONAL EN FIN DE TROISIEME.

Cet examen concernerait obligatoirement tous les élèves scolarisés à ce niveau. Il devrait remplir mieux qu'aujourd'hui les trois fonctions définies au début de notre propos :

- **Diplôme national de fin d'études obligatoires, il garantirait l'acquisition des savoirs et des compétences nécessaires à la bonne intégration sociale, professionnelle et personnelle de « l'honnête citoyen » de notre siècle.**

- **Il garantirait, dans chaque discipline enseignée au collège, l'acquisition des « savoirs et des compétences transversales de base » nécessaires au succès d'études ultérieures diversifiées (technologique ou professionnelle aussi bien que générale).**

- **Il serait susceptible de fournir à l'institution (aux différents niveaux) les informations et les outils nécessaires à l'analyse des résultats des élèves et au pilotage des établissements.**

Ce diplôme devrait retrouver son caractère de **diplôme national** et cesser d'apparaître comme un examen « maison ». Dans cet esprit, nous proposons que les séries technologiques (dès lors que les classes de 3ème technologiques n'existent plus) et professionnelles, soient supprimées. **L'examen serait réservé exclusivement aux élèves des classes de 3ème des collèges.** Les élèves qui n'en disposent pas en fin de classe de 3ème pourraient s'y présenter l'année suivante en candidats libres.

5.2 - REDONNER A L'EXAMEN SA VALEUR ET SA DIGNITE

- **Les sujets des épreuves terminales**

Nous l'avons vu, l'Inspection générale de l'Education nationale, a entrepris un travail important depuis 1999 pour que les sujets :

- répondent mieux aux exigences du système éducatif ;
- soient en phase avec les programmes du collège de la 6ème à la 3ème et plus spécifiquement pour les éléments qui la concernent, ceux de la classe de 3ème ;

- garantissent les connaissances de base dans chacune des disciplines, en fin de scolarité obligatoire, étant bien entendu que les éléments de cette culture de base de « l'honnête homme du XXIème siècle » devrait faire l'objet d'un arbitrage politique clair et courageux.

Il n'est pas question de vouloir rétablir un examen centralisé à sujets nationaux, mais le choix des sujets devrait être systématiquement piloté par l'Inspection générale, dans chacune des disciplines du contrôle terminal. Nous proposons que les sujets académiques (ou inter-académiques) soient examinés et validés régulièrement par l'Inspection générale.

On pourrait utiliser efficacement les travaux menés dans le cadre des évaluations nationales qui ont conduit à créer un système de banque d'exercices qui sera progressivement mise en ligne. Les inspecteurs et professeurs qui conçoivent les sujets pourraient disposer ainsi d'exemples d'exercices garantissant une meilleure cohérence inter-académique. Cette méthode pourrait conduire à une meilleure harmonisation des corrections.

Le Brevet, nous l'avons vu, n'est pas vraiment « piloté » pédagogiquement. Il devrait l'être, désormais, au niveau académique (le pilotage administratif et l'organisation de l'examen au plan départemental donnent toute satisfaction et seraient donc, maintenus). Ce pilotage pédagogique porterait sur le choix des sujets, les barèmes, l'harmonisation de la notation (en amont et en aval de la correction), la composition des jurys, une fixation académique des normes de repêchage, la remontée des résultats et leur traitement annuel.

L'utilisation des résultats pour le pilotage et le suivi des établissements (comparaison de l'évaluation en 6ème et des résultats au Brevet : suivi de cohorte, analyse de « la valeur ajoutée » de chaque collège) sont menés dans certains départements. Elle serait généralisée et pour cela, la DPD pourrait fournir aux Recteurs et aux Inspecteurs d'Académie des outils communs permettant, non seulement l'analyse locale des résultats, mais aussi les comparaisons départementales, académiques voire nationales et leur confrontation régulière avec les résultats des évaluations nationales de fin de 3ème.

Par ailleurs il semble difficile – nous l'avons vu – que ne figurent pas, parmi les épreuves terminales, une épreuve en langue vivante et une en informatique (en informatique il suffirait sans doute, d'utiliser le Brevet informatique qui vient d'être créé). Ceci aboutirait inévitablement à un renforcement et à une généralisation de l'enseignement de l'informatique, en assurant sa « démocratisation ».

- **Le contrôle continu**

La mise en œuvre du contrôle continu, dans l'attribution du Brevet semble une excellente formule qui a fait ses preuves dans d'autres examens. La plupart des professeurs y sont très attachés. Un travail d'encadrement de même nature que celui prévu pour les épreuves écrites devrait être engagé.

Dans un premier temps, il suffirait de conduire les établissements à appliquer les dispositions prévues par les textes de 99.

Dans un second temps, il faudrait, qu'en appui aux corps d'inspection, des outils communs soient mis à la disposition des académies par la DPD et l'Inspection générale, afin d'homogénéiser les pratiques et permettre les comparaisons. Il ne saurait être question pour nous de priver les professeurs de leur responsabilité dans la notation et l'évaluation de leurs élèves ni de remettre

en cause le principe selon lequel « celui qui enseigne assume la responsabilité de noter et d'évaluer », mais de donner des cadres, des outils et des règles communes.

La formation (obligatoire) des professeurs à la pratique du contrôle continu devrait être lancée rapidement dans le cadre des plans académiques de formation continue, à l'instar de ce qui a été fait dans l'enseignement technique.

Nous proposons enfin que les disciplines qui font l'objet d'un contrôle terminal soient, toutes, évaluées systématiquement dans le cadre du contrôle continu, en classe de 4ème comme en classe de 3ème : l'intervention de professeurs différents constitue – on le sait – une garantie supplémentaire de validité.

- **Le brevet et l'orientation des élèves**

Nous l'avons vu, pour toute une série de raisons historiques, les résultats des élèves au Brevet ne sont pas pris en compte dans l'orientation des élèves. Le rapport de 1991 de l'Inspection générale relevait cette anomalie, en s'étonnant que les notes aux épreuves terminales qui, seules, lui semblaient significatives du niveau de l'élève, ne soient pas prises en compte.

Nous proposons que ces résultats soient partie intégrante du dossier d'orientation de l'élève et qu'ils soient transmis, avec l'ensemble du dossier, aux professeurs principaux des établissements d'accueil (comme le dossier d'entrée en 6ème). On objectera des problèmes de calendrier, car les épreuves terminales du Brevet se déroulent en toute fin d'année, après les procédures d'orientation : il suffirait d'inverser l'ordre des opérations en organisant les épreuves terminales au début du 3ème trimestre, et d'arrêter les notes de contrôle continu en fin d'année, pour permettre cette prise en compte. Précisons **qu'il ne s'agit nullement de transformer les épreuves terminales en « examen de passage » en classe de seconde,** ce qui constituerait une aberration par rapport aux objectifs des réformes successives et à la volonté affichée de démocratisation de l'enseignement secondaire. Nous ne pouvons imaginer – comme certains esprits pessimistes – que les épreuves terminales étant passées au début du 3ème trimestre, les élèves quitteraient prématurément leurs établissements. D'ailleurs les notes de contrôle continu ne seraient arrêtées qu'en toute fin d'année et les professeurs pourraient dans ce cadre, organiser les ultimes épreuves communes prévues par les textes de 1999.

5.3 - GARANTIR QUE LES TITULAIRES DU BREVET DISPOSENT BIEN DES SAVOIRS DE BASE

En plus des mesures proposées, il nous semble que certaines innovations pourraient être tentées :

- **Parmi les épreuves de contrôle terminal, ne serait-il pas envisageable de créer, en liaison avec la DPD, une épreuve standardisée de niveau, permettant les comparaisons individuelles et temporelles ?**

En plus des évaluations nationales périodiques qui font l'objet d'autres propositions, nous disposerions d'un élément supplémentaire annuel susceptible de bien éclairer les enseignants sur le niveau de leurs élèves, et les responsables du système éducatif sur l'évolution des compétences des jeunes. En outre cette épreuve pourrait fournir d'utiles informations aux établissements d'accueil des élèves.

Pour ce faire il serait nécessaire de répondre d'abord aux deux questions suivantes :
- souhaite-t-on un effet d'ancrage, la moyenne aux protocoles permettant de corriger les

moyennes des classes, des établissements ?
- s'agit-il d'introduire un résultat supplémentaire et quel poids souhaite-t-on lui attribuer ?

Ces épreuves seraient progressivement intégrées dans le contrôle terminal. Elles permettraient de situer l'ensemble des élèves sur une même échelle et de mettre en œuvre des comparaisons temporelles intéressantes. Le détail de nos propositions apparaît en annexe 4.

- **Les capacités d'expression et de communication des élèves ainsi que leur aptitude à développer une argumentation devraient être mieux évaluées.**

Pour cela, pourrait figurer dans le cadre du contrôle continu, la note obtenue à la soutenance devant quelques professeurs, d'un mémoire, rédigé pendant les deux années de 4ème et 3ème, à partir des activités de travaux personnels encadrés, de thèmes transversaux ou des parcours diversifiés.

Pour autant, cette réforme ne devrait pas conduire à durcir exagérément le niveau de l'examen en réservant le succès aux élèves les plus brillants. Si le diplôme garantit que tous les titulaires disposent – en fin de scolarité obligatoire – des connaissances de base requises, les élèves sont, néanmoins, de niveau très différent, dans chaque établissement et d'un secteur à l'autre. Pour mieux rendre compte de cette diversité, nous proposons de **distribuer des mentions comme au baccalauréat : la mention passable garantirait – comme dans tout diplôme national – que le niveau de base est atteint.**

Bien entendu un système d'options serait proposé, correspondant à la diversité des parcours ou des choix individuels : « itinéraires de découverte » du plan ministériel « Pour un collège républicain » : technologie, informatique, arts, éducation physique, etc.

Il nous semble enfin, qu'il serait intéressant de lier dans cet examen la certification et le bilan individuel de compétences. Pour cela on pourrait adopter un système de niveaux dans chacune des disciplines figurant au contrôle terminal. Ce système en niveaux de compétences serait sans doute mieux accordé aux exigences du monde du travail et il permettrait plus facilement l'insertion de l'adulte dans la politique de « formation tout au long de la vie » qui finira bien par se mettre en place un jour et se développer (voir annexe 4).

II LES EVALUATIONS - BILANS DE LA DEP/DPD

Rapport conduit par Pierre VRIGNAUD
avec la participation de Denis BONORA
et les contributions de Philippe CHARTIER et Bruno TROSEILLE,
Service de Recherche
de l'Institut National d'Etude du Travail et d'Orientation Professionnelle,
Conservatoire National des Arts et Métiers.

INTRODUCTION : LES DIFFERENTS TYPES D'EVALUATION

Quelles informations fournissent les évaluations nationales réalisées par le MEN, et en particulier celles de la DPD (Direction de la Programmation et du Développement), sur les acquis des élèves en fin de collège ? Comment améliorer les dispositifs d'évaluation afin de mieux répondre aux besoins des différents utilisateurs ?

Nous traiterons ces différents aspects en nous attachant dans un premier temps à identifier les informations apportées par ces dispositifs, à définir leurs qualités psychométriques, et à examiner les types de traitement auxquels ils ont donné lieu, et les interprétations qu'ils permettent. Nous discuterons des possibilités de leur amélioration. Deux questions feront l'objet d'une attention particulière : comment étendre les domaines évalués ; comment satisfaire au mieux les exigences de l'étude des perspectives comparatives, temporelle et transversale.

L'une des questions centrales qui conditionnent la mise en œuvre d'un dispositif national d'évaluation est celle du choix de son orientation générale, qui devrait être précisée avec netteté dès la phase initiale, à défaut de quoi les informations recueillies risqueraient d'être en bonne partie inexploitables. Il est fréquent en effet que les données collectées sous une orientation soient peu compatibles avec un traitement utilisable dans une autre perspective.

Parmi les différentes visées d'évaluation envisageables au plan national (ou régional) on distingue d'abord celle qui porte sur les variables « d'entrée », qu'on peut appeler les « circonstances » de l'enseignement (environnement matériel, programmes, types et « quantité » de formation des maîtres, nombre de professeurs et d'élèves, carte scolaire, transport scolaire, etc…). En France, ces études fondamentales, qui débouchent sur un grand nombre de constats en terme de dénombrement, flux, etc., sont réalisés par le MEN/DPD et font l'objet d'intéressantes publications annuelles.

Les informations sur ces variables d'entrée peuvent être réinvesties dans le cadre de certaines études se situant dans le champ d'une autre visée d'évaluation, qui nous occupe ici plus directement.

Il s'agit de l'évaluation des acquisitions des élèves (connaissances, attitudes, capacités psychomotrices, etc…). A l'intérieur de ce champ il nous faut d'abord distinguer une perspective axée sur **les élèves, individuellement**, et une autre axée sur **la populatio**n (ou des sous-

populations) en vue de tirer des conclusions sur le fonctionnement du système[1].

Dans la perspective **axée sur les élèves individuels**, l'intention peut être certificative ; elle peut aussi être diagnostique.

Au niveau du collège, où nous nous situons dans ce rapport, l'évaluation **certificative** permet d'officialiser une reconnaissance des acquis de chaque élève au cours de sa scolarité (brevet). Les contraintes sont celles inhérentes à tout examen débouchant sur la délivrance d'un diplôme. Un problème qui se pose ici, parmi d'autres, est celui de l'introduction d'une procédure qui permette de garantir l'équivalence - au moins approximative - d'un diplôme obtenu, quel que soit le lieu géographique de la passation de l'examen.

Quant à l'évaluation **diagnostique**, elle devrait permettre aux maîtres de suivre la progression de leurs élèves, et à tout le moins, d'identifier leurs lacunes en vue d'y remédier. Dans ce cadre, on imagine aisément que, même si la référence pour la construction de l'instrument d'évaluation est nationale - à savoir les programmes officiels -, le choix des items d'évaluation devra tenir compte, pour ce qui concerne les connaissances par exemple, des aspects du fonctionnement psychologique de l'élève au cours de sa progression vers l'objectif pédagogique. Pour chaque objectif, la hiérarchie des apprentissages devra être jalonnée, chaque jalon devant être représenté par des items, construits pour évaluer le niveau de maîtrise atteint par l'élève (évaluation critérielle). Le psychopédagogue et le psychologue pourront dans ce contexte apporter une contribution déterminante.

Dans la perspective **axée sur la population**, on s'intéresse au fonctionnement du système, et dans ce cadre on cherche à établir un bilan par rapport aux objectifs terminaux d'un cycle d'étude ou d'un niveau dans le cursus. Les observations recueillies devraient alors permettre d'évaluer le rendement du système, notamment par la comparaison avec le niveau atteint antérieurement dans le cursus, et de définir l'importance relative des différentes variables d'entrée (évoquées plus haut) qui contribuent à la formation scolaire. Le constat éventuel d'une hétérogénéité de niveau entre sous-groupes de sujets devrait également conduire à étudier de façon différentielle l'effet de ces variables d'entrée, pour aboutir à des préconisations ciblées sur les différentes sous-populations concernées. Les comparaisons temporelles et entre systèmes éducatifs s'intègrent également dans ce cadre. Ces différentes approches paraissent indispensables à un pilotage efficace du système (on peut noter que cette conception de l'évaluation répond dans son principe à un modèle technocratique. Poussée à l'extrême, elle risquerait de stériliser les initiatives locales. Il est toutefois possible de concevoir à l'inverse qu'elle soit compatible avec une certaine dose d'autonomie des établissements - sans doute souhaitable - dans la mesure où les objectifs évalués pourraient être circonscrits aux apprentissages jugés indispensables pour garantir la formation scolaire considérée).

Or, les contraintes auxquelles devra satisfaire l'élaboration d'un dispositif d'évaluation seront largement différentes selon que celui-ci est prévu pour remplir une fonction diagnostique ou une fonction de bilan (on trouvera un aperçu de cette hétérogénéité des contraintes en annexe 6). De ce fait ces visées sont peu compatibles entre elles, et peuvent difficilement coexister dans un même dispositif.

Dans cette partie du rapport, c'est le type d'évaluation à visée de bilan, tel qu'il est mis en pratique par les services du MEN, qui sera considéré en principal. Nous présenterons d'abord les

[1] Un éclairage complémentaire, basé sur les différentes fonctions de l'évaluation, est proposé dans Bonora (1996),ainsi que dans l'annexe 6.

constats puis nous formulerons différentes propositions pour améliorer les dispositifs.

1 - LES CONSTATS

1.1 - PRESENTATION DE L'ETUDE

1.1.1 - Méthodologie

Nous avons réalisé cette étude à partir :

- d'une analyse de l'ensemble des documents des évaluations-bilans niveau troisième : protocoles, documents de présentation des résultats publiés ou à usage interne (cités dans les références) ;
- d'une analyse approfondie des protocoles dans deux disciplines : histoire/géographie et physique/chimie (voir annexes 7 et 8) ;
- des résultats d'analyses psychométriques sur les protocoles de différentes disciplines et de plusieurs années ;
- des rapports d'expertises effectuées antérieurement par l'INETOP sur les comparaisons temporelles en mathématiques ainsi que sur les banques d'outils ;
- d'entretiens avec les responsables et personnels de la Mission d'évaluation des élèves ;
- d'une enquête par courrier électronique auprès des organismes travaillant sur l'évaluation des acquis des élèves de différents pays européens ;
- d'une visite auprès de deux organismes étrangers : CITO aux Pays-Bas, *Landesinstitut für Schule und Weiterbildung* du Land de Nordrhein-Westfalen en République Fédérale d'Allemagne.

1.1.2 - Les évaluations de la DPD

La Mission de l'évaluation des élèves de la DPD (et les services qui l'ont précédée : SIGES, SPRESE, DEP) conduit des enquêtes sur l'évaluation des acquis des élèves. Il n'entre pas dans notre propos de détailler les dispositifs de la DPD, mais pour bien situer notre travail, de rappeler ses deux principaux types de dispositifs systématiques. Nous nous intéressons aux connaissances des élèves en fin de troisième. Ces informations sont produites par **les évaluations-bilans** qui ont lieu périodiquement (environ tous les cinq ans depuis 1984). Les évaluations de type bilan sont conduites à différents niveaux de notre système éducatif, notamment en fin de troisième ; elles ont pour *« …finalité première … de fonder le pilotage du système éducatif sur les résultats pédagogiques de son fonctionnement : il s'agit de mieux connaître la réalité présente pour mieux préparer l'avenir. »,* dispositif 1984, document de présentation des résultats, p. X.

Il s'agit d'enquêtes visant à obtenir des résultats au niveau de la population à partir d'un échantillon représentatif (tiré a priori).

Il ne faut donc pas les confondre avec les **évaluations à finalité diagnostique** et formative qui sont proposées à l'ensemble des élèves aux niveaux CE2, sixième, seconde. Ces dernières ont pour but de donner aux enseignants et aux équipes éducatives des informations sur les acquis des élèves, afin de mieux gérer les apprentissages dans la classe. Une certaine confusion peut s'établir entre ces types d'évaluation dans la mesure où la DPD publie des résultats nationaux en utilisant les résultats obtenus à ces évaluations diagnostiques pour un échantillon tiré a posteriori.

Notre étude porte principalement sur les évaluations de type bilan, ou évaluations sommatives, et non sur les évaluations diagnostiques, qui d'ailleurs ne sont pas réalisées au niveau de la troisième. Nous avons évoqué, plus haut, les différences entre évaluations diagnostique d'une part, et sommative de l'autre. Or, l'examen de l'évolution du dispositif et de son mode de construction nous a montré que s'étaient sans doute produites des contaminations entre les représentations des différents dispositifs. Ces contaminations ont eu des effets préjudiciables sur la fiabilité du dispositif des évaluations nationales et de la lecture de leurs résultats.

1.1.3 - Bref historique

A notre connaissance, la première grande enquête française sur les acquis des élèves au niveau troisième est la recherche conduite en 1963/1964 par l'INETOP sur l'orientation à la fin du premier cycle secondaire (Reuchlin et Bacher, 1969). Cette recherche, qui portait principalement sur les déterminants de l'orientation, comprenait deux épreuves standardisées d'évaluation des connaissances (français et mathématiques). Par la suite, différentes enquêtes ont été conduites dans les années 1970 pour aboutir aux enquêtes nationales sur les acquis des élèves en primaire (fin 1970) et au collège (1980 pour les sixièmes, 1982 pour les cinquièmes). Dans le prolongement de cette dernière enquête, a été mis en place, en 1984, le premier dispositif portant sur les acquis des élèves de 3éme par le Service de la Prévision, des Statistiques et de l'Evaluation (SPRESE). L'évaluation au niveau troisième a été reconduite (avec des protocoles plus ou moins fortement modifiés et en incluant différentes disciplines) en 1988 (pour les seules troisièmes technologiques), 1990 (Direction de l'Evaluation et de la Prospective-DEP), 1995 (DEP) et enfin 1999 (Direction de la Programmation et du Développement-DPD). On trouvera dans les références le détail des disciplines évaluées pour chacune de ces évaluations.

1.2 - QUE MESURENT LES PROTOCOLES D'EVALUATION ?

Rappelons que les protocoles sont des épreuves standardisées regroupant des exercices. Chaque exercice comprend en général plusieurs questions. Il est d'usage dans le champ de la mesure de considérer ces questions comme des items. Les questions se présentent sous différents formats : à choix multiples (QCM), ouvertes (réponse courte, justification d'une démarche). On rencontre également des situations de production de textes. Ces dernières situations se distinguent, par exemple, de la dissertation, par la mise en place d'une standardisation de la situation de production, ainsi que du codage utilisé par les correcteurs. Rappelons également que ces protocoles sont administrés selon un calendrier identique pour tout l'échantillon.

1.2.1 - Objectifs pris en compte par les évaluations de type bilan

Les évaluations de type bilan visent à déterminer dans quelle mesure les objectifs fixés par les programmes sont atteints. Les informations collectées sur les performances des élèves s'inscrivent dans une conception des connaissances et des compétences de type scolaire. Cette intention est manifestée par le rappel des programmes au début des publications présentant les résultats. Les exercices et les items sont élaborés à partir d'une nomenclature d'objectifs identifiés à partir des programmes. Cette conception est, de plus, sous-tendue par les principes de construction des épreuves. Les protocoles sont construits par des groupes réunissant les représentants des principaux acteurs de l'enseignement dans chaque discipline (Inspection générale, professeurs) et des personnels de la DPD.

Une expertise sur la pertinence des exercices est ensuite obtenue en demandant, lors des

passations, l'avis des professeurs sur l'importance de l'objectif que chaque exercice évalue. Cette manière de procéder, tout à fait classique en ce qui concerne des protocoles d'évaluation des élèves, permet de vérifier la couverture du programme et la représentativité des exercices et des items.

Les objectifs sont hiérarchisés selon une taxonomie en trois niveaux, depuis les objectifs les plus généraux jusqu'aux items (voir les annexes 7 et 8 qui présentent l'analyse des objectifs de deux disciplines). Un des principaux problèmes vient du caractère trop descriptif des objectifs les plus fins. Ces taxonomies sont certainement signifiantes, mais plutôt pour des enseignants de la discipline concernée. Les objectifs du niveau hiérarchique le plus élevé semblent davantage accessibles à un public de non-spécialistes. La terminologie utilisée supporte mal, selon nous, le transfert vers des utilisations plus générales : pilotage du système éducatif dans son ensemble, rapprochement avec le domaine des compétences professionnelles pour une évaluation tout au long de la vie. On peut constater l'écart entre les définitions des compétences en comparant par exemple les compétences en langues vivantes, telles qu'elles sont définies dans les évaluations DPD, et les compétences en langues vivantes définies dans le cadre d'épreuves de langues vivantes à usage professionnel (TOEFL - *Test Of English as Foreign Language*, passeports linguistiques du conseil de l'Europe). Par ailleurs, cette dépendance à l'égard des programmes peut poser problème pour les comparaisons temporelles. La terminologie a subi au fil des évaluations des modifications. Si l'on évalue toujours la même chose au regard des exercices et des items dans certaines disciplines, dans d'autres les objectifs et leur dénomination ont changé.

La méthode employée vise à évaluer principalement les acquis correspondant aux objectifs pédagogiques de la classe de troisième ou du cycle dont elle fait partie. Certes, les protocoles comprennent des exercices et des items qui correspondent à des objectifs antérieurs à la troisième voire à l'entrée au collège. Mais, il nous paraît difficile de considérer que les évaluations en classe de troisième puissent, en l'état, informer sur les connaissances acquises au cours de la scolarité obligatoire. Un tel objectif nécessiterait de repenser les principes de construction des protocoles. Notons cependant que la terminologie évolue vers des objectifs définis de manière plus indépendante des programmes, lorsque les analyses s'ouvrent sur des problématiques de comparaisons temporelles ou internationales.

Un dernier point qui montre la prépondérance de la perspective scolaire, est la présence majoritaire des disciplines jugées fondamentales et la moindre présence ou l'absence totale des disciplines dites mineures (artistiques en particulier). Certes, l'évaluation standardisée des matières artistiques est difficile, mais elle s'avère réalisable comme le montre une étude pilote réalisée par la DPD (Levasseur & Shu, 1998). Ces disciplines font partie des dispositifs d'évaluation dans plusieurs des pays sur lesquels nous avons recueilli des informations. Et pourquoi l'éducation physique n'est-elle pas intégrée alors que son mode d'évaluation se prête tout à fait à la standardisation. On remarquera de même que certaines compétences, pourtant souvent citées comme indispensables à la vie professionnelle, sont habituellement absentes des objectifs d'évaluation : la communication orale et la maîtrise des nouvelles technologies de l'information et de la communication. En ce qui concerne la communication orale, une expérimentation a été faite sur la classe de seconde (Levasseur et Trussy, 1997). Celle-ci a montré que ce type d'évaluation pouvait être mis en place malgré la lourdeur de la mise en œuvre. L'évaluation des nouvelles technologies de l'information et de la communication (TICE) est tout à fait envisageable. C'est ainsi, par exemple, que l'Inspection Générale de l'Education Nationale a mené une évaluation pilote dans une académie à l'aide d'une épreuve sur ordinateur mise à disposition par le CNDP (IGEN, 2000 et 2001).

1.2.2 - Autres objectifs pris en compte

Notre revue de questions à partir des publications de la DPD nous a permis de constater qu'il existait parmi les dispositifs d'évaluation de la DPD ou dans d'autres opérations (en particulier des évaluations internationales) des informations de nature différente ou plus larges que les connaissances scolaires définies à partir des programmes. Signalons par exemple une enquête sur la production écrite dans des pays francophones qui présente des vues intéressantes sur l'identification des compétences pour la production de textes et la manière de les évaluer. Citons également les études basées sur les réponses aux questionnaires "Vie Scolaire" et « Méthodes de travail » inclus dans les évaluations de type bilan qui donnent des informations sur des compétences transversales telles que les méthodes de travail, l'éducation à la citoyenneté. Citons enfin l'étude sur les acquis socio-affectifs (Grisay, 1997a) qui balaie un ensemble de compétences « sociales », ne faisant pas partie explicitement des programmes disciplinaires, mais pouvant être considérées comme des objectifs implicites de l'éducation.

1.2.3 - Relations avec le brevet

Il existe une seule étude publiée sur ce thème, qui met en relation les résultats des échantillons d'élèves ayant passé l'évaluation de 1995 avec leurs résultats au brevet des collèges (Murat, 1998). Cette étude est très soignée sur le plan statistique (on regrettera cependant, l'absence d'étude psychométrique préalable). La relation entre les résultats aux protocoles des évaluations-bilans et au brevet peut être considérée comme relativement élevée compte tenu des multiples différences dans les conditions de réalisation (sujets, format, durée, mode de correction, motivation, etc.). La force de la liaison estimée par la part de variation commune aux deux évaluations est comprise entre 25 et 50% selon les disciplines. Elle est plus élevée dans les disciplines scientifiques que littéraires, et faible en technologie. On peut penser que cette variation est liée en partie à la plus ou moins grande facilité d'évaluer les objectifs de la discipline par des épreuves standardisées. L'importance de cette liaison ne doit pas donner l'illusion que les notations sont totalement fiables et ne présentent pas de variabilité selon les contextes. Les liaisons entre les différentes évaluations indiquent uniquement une relation entre les classements des élèves. Les élèves sont en gros ordonnés de la même manière par les épreuves standardisées et par les notes à l'examen et dans une moindre mesure au contrôle continu. Il existe par contre des disparités importantes entre les échelles de notation entre les classes ce que l'on appelle un « effet-classe »[2]. Selon la classe dans laquelle ils sont scolarisés, les élèves peuvent être systématiquement sur-notés et sous-notés. Ces analyses retrouvent les grandes conclusions des études docimologiques sur les biais de la notation, réalisées depuis un demi-siècle (Piéron, 1963).

Indépendamment de l'intensité de la liaison entre les mesures, le brevet, sous sa forme actuelle fournit peu d'informations sur les connaissances des élèves - est-ce réellement son objectif ? - du fait du caractère à la fois global et disparate de la note qui rend compte des performances à des épreuves de contenu très différent d'une académie à l'autre, et d'un établissement, voire d'un enseignant à l'autre.

1.3 - ETUDE DE FIABILITE DES PROTOCOLES DE LA DPD

1.3.1 - Importance des analyses psychométriques

[2] Il serait intéressant d'employer les modèles multiniveaux bien adaptés pour ce type d'analyses (voir Jarousse & Leroy-Audoin, 2001).

L'étude de nos savoirs sur les connaissances des élèves en fin de collège conduit à répondre à deux questions : ce que nous savons et comment nous le savons. La seconde question est préalable à la première car la manière dont nous recueillons les informations conditionne leur nature et leur fiabilité. Nous avons souvent trouvé dans les publications de la DEP des annexes, des encarts méthodologiques très pertinents et didactiques, qui témoignent du souci apporté à la fiabilité statistique du dispositif et à la communication de réserves éventuelles des statisticiens. Cependant, nous avons constaté l'absence quasi-totale de références aux concepts et méthodes de la mesure en éducation, de la psychométrie (il ne nous semble pas avoir lu une seule fois le mot dans l'ensemble des documents concernant les évaluations de fin de troisième cités en références).

Nous allons présenter très brièvement l'esprit de la démarche psychométrique ; pour une présentation plus complète, nous renvoyons le lecteur à des ouvrages généraux sur la mesure en éducation comme Linn, 1989 ; de Landsheere, 1994 ; ou sur la psychométrie, Dickes et al., 1994. Schématiquement, l'objet de la psychométrie est de construire une mesure fiable. En psychométrie, on considère que le score observé est composé d'au moins deux parts : le score vrai et l'erreur de mesure. Le concept d'erreur en psychométrie intervient à tous les stades de la démarche : du choix des questions jusqu'à l'interprétation des résultats. Les études de validation ont pour but d'identifier et de réduire autant que faire se peut les différentes sources de biais pour garantir la fiabilité des inférences finales sur l'interprétation des résultats.

La distinction entre performance et compétence est une autre manière de poser le problème de la mesure en éducation (sur cette distinction voir en particulier Flieller, 2001). On observe les performances des élèves pour conclure sur leurs compétences. Les compétences ne sont pas directement observées. On va inférer sur les compétences à partir des performances. Les performances sont dépendantes de nombreux autres facteurs que de la seule compétence évaluée, facteurs qu'il faudra essayer de prendre en compte.

1.3.2 - Analyse des protocoles de la DPD

Abordons maintenant les conclusions de nos analyses réalisées aux différentes étapes d'élaboration des protocoles de la DPD.

1.3.2.1 - L'échantillonnage des items

La première étape est l'échantillonnage de l'univers des items. Ce point est en général moins bien connu que l'échantillonnage des sujets. Il est spécifique des problèmes de la mesure en sciences humaines en général et en éducation en particulier. En effet, il est nécessaire que le protocole soit représentatif de l'univers à évaluer. On concevrait mal un protocole de mathématiques qui ne comporterait que des exercices d'algèbre et aucun exercice de géométrie. De même que la représentativité de l'échantillon des sujets permet de généraliser sur la population parente (non observée), de même la représentativité de l'échantillon des items quant au domaine évalué permet de généraliser sur l'ensemble du programme (non observé).

La méthode mise en œuvre pour la construction des protocoles de la DEP remplit en principe cette fonction. Le programme est découpé en objectifs ; chaque objectif est éventuellement pondéré en fonction de son importance dans le programme. Cependant, le nombre d'items est plus ou moins proportionnel à l'importance de l'objectif et, parfois même, certains objectifs ne sont pas représentés du tout.

Cependant, il existe d'autres méthodes pour s'assurer de cette représentativité. Signalons par exemple la méthode de Bloom, basée sur sa taxonomie. Une approche intéressante a été menée par Alain Lieury (1996) et son équipe dans leur étude de l'acquisition et de l'organisation des connaissances en mémoire chez les élèves de collège. Ils ont identifié, dans les manuels scolaires, les termes nouveaux qui apparaissaient pour chaque discipline à chacun des niveaux scolaires de la sixième à la troisième. Ils ont ensuite construit des questionnaires en échantillonnant sur l'ensemble des termes identifiés pour un niveau donné. Cette manière de procéder fournit une autre approche pour tenter de construire « objectivement » la représentativité de l'échantillon de questions face à l'univers des connaissances à évaluer au collège.

1.3.2.2 - La standardisation

Les conditions de standardisation des passations sont bien prises en considération. On peut néanmoins s'interroger sur un certain nombre d'éléments implicites susceptibles de compromettre cette standardisation. Signalons par exemple la présence éventuelle de différences de conditions de passation entre les classes, et entre les établissements. Ainsi, dans les cas où les élèves passaient deux cahiers d'épreuves, le délai entre les deux passations pouvait varier : soit que les élèves passaient les deux cahiers dans une matinée, soit qu'ils en passaient un le matin, l'autre l'après-midi. On peut penser que dans la première condition, les effets de fatigue et de lassitude étaient plus importants que dans la seconde.

Nous n'avons pas trouvé d'études sur le respect de la standardisation du codage par les correcteurs (fidélité inter-juges). Ces vérifications sont indispensables. Elles paraissent d'autant plus nécessaires que les protocoles comportent souvent des questions ouvertes à réponses longues ainsi que la production de textes. Dans ce dernier cas, le dispositif devrait comporter une formation des correcteurs pour garantir une harmonisation maximale.

Par ailleurs, les dates de passation en fin d'année de troisième, et le fait que la performance à cette épreuve n'entre pas dans la notation de l'élève, sont autant de facteurs de démotivation qui pourraient réduire la qualité de la performance.

1.3.2.3 - Analyse interne des épreuves

En général, les analyses internes se basent sur trois types d'indicateurs : deux au niveau de l'item : sa difficulté (la fréquence de réussite à l'item), sa discrimination (la liaison entre l'item et le score total) ; et un au niveau du score ou des sous-scores : un indicateur permettant d'apprécier la consistance de l'ensemble des items, c'est-à-dire le fait que l'ensemble des items appartiennent à un même domaine, homogène, et que le score reflète bien la compétence dans le domaine évalué.

1.3.2.3.1 - La difficulté des item

Dans l'approche psychométrique classique, on désigne sous le terme de difficulté des items la fréquence de réussite (pourcentage de bonnes réponses).

Les difficultés des items sont utilisées pour piloter le niveau de difficulté global de l'épreuve. Il faut en effet se garder de penser que les avis des experts (e.g. les enseignants) sont suffisants pour évaluer la difficulté de l'item. La seule mesure objective est la fréquence de réussite dans un échantillon représentatif. La lecture des résultats de l'évaluation de 1984 où sont mises en relation difficulté estimée par les professeurs et fréquence de réussite, montre bien souvent des

écarts importants (d'au moins 10% en moyenne et souvent de 20 à 30 %).

Il faut également rappeler que la difficulté d'un item n'est pas une mesure absolue, elle est dépendante de l'échantillon. La difficulté d'un protocole est liée aux intentions des constructeurs. Nous n'avons pas trouvé d'indications sur les intentions des constructeurs quant aux niveaux de difficulté recherchés. Nous avons constaté des différences entre les années, les épreuves de 1984 étant en moyenne plus difficiles (50% d'items réussis en moyenne) que les épreuves des années postérieures (75%).

Outre l'image qu'il donne de la maîtrise des objectifs, le niveau global de difficulté d'un protocole a plusieurs implications méthodologiques importantes. En effet, le niveau de difficulté moyen dépend de la distribution des scores. En psychométrie, on cherche souvent à éliminer les items dont la fréquence de réussite est trop élevée (e.g. > 90%, effet plafond) ou trop faible (e.g. < 10%, effet plancher). Un protocole facile (réussi à plus de 80% en moyenne) induit une distribution asymétrique, c'est à dire que les résultats des sujets les plus performants sont moins dispersés que les résultats des sujets faibles et moyens. Un tel protocole est, de ce fait, peu sensible pour distinguer les sujets à performance élevée des sujets à performance très élevée. Un tel protocole, par contre, distinguera bien les sujets faibles des sujets moyens. La distribution des scores a d'autres conséquences, car elle joue sur les liaisons entre le score et d'autres indicateurs qui seront étudiés : par exemple les relations entre les disciplines. De plus, dans la perspective d'étude des évolutions temporelles, la marge d'évolution de protocoles faciles, est plus restreinte, car le tassement observé à la première passation aura tendance à s'accentuer à la seconde.

En conclusion, nous dirons que si le choix du niveau de difficulté appartient aux constructeurs du protocole, il est par contre indispensable d'expliciter les raisons de ce choix.

1.3.2.3.2 Discrimination et consistance interne

Ni les indicateurs de discrimination[3], ni les indicateurs de consistance n'ont été publiés. Quelle est la conséquence de cette absence ? On peut penser que si des analyses ont été faites lors des pré-expérimentations (mais cela n'a pas toujours été le cas), ces analyses ont permis de s'assurer de la consistance interne des protocoles et de la qualité des items. Cependant, les analyses que nous avons effectuées sur des protocoles de différentes disciplines et de différentes années montrent que cela n'est pas toujours vrai. Il existe dans la plupart des épreuves une proportion d'items présentant une discrimination faible voire inverse (cf. note précédente). Les protocoles de certaines matières (français, histoire géographie) apparaissent globalement peu consistants et comportent une proportion importante d'items de mauvaise qualité (plus de 10% et jusqu'à 20% pour certaines épreuves de 1984 que nous avons analysées). Le score global de ces protocoles est fortement biaisé. Les inférences à partir de ce score sont, de ce fait, d'une fiabilité réduite. Ce problème se répercute évidemment sur l'ensemble des analyses utilisant ce score - par exemple les comparaisons selon le genre, les PCS.

[3] Le concept de discrimination peut apparaître moins habituel que celui de difficulté. De manière générale, il répond à la question « les sujets qui réussissent l'item ont-ils une compétence plus élevée que les élèves qui échouent ? ». Cette question est opérationnalisée en comparant les performances au test des sujets qui réussissent et des sujets qui échouent. On s'attend à ce que le score des sujets qui réussissent soit en moyenne plus élevé que celui des sujets qui échouent. Si c'est le cas, alors on peut dire que l'item discrimine bien les sujets forts (du point de vue de la compétence) des sujets faibles. On voit l'intérêt de cette notion : un item présentant une bonne discrimination apporte une information importante sur les compétences des sujets ; un item de discrimination faible apporte peu d'information, il n'est pas très utile du point de vue du protocole ; un item de discrimination nulle, n'apporte pas d'information voire introduit du bruit dans la mesure ; un item de discrimination inverse (les sujets qui le réussissent posséderaient moins la compétence que les sujets qui échouent !) fonctionne à l'envers.

1.3.2.4 - Problèmes posés par la généralisation sur un objectif à partir d'un nombre restreint d'items

La généralisation sur un objectif à partir d'un item pose problème du fait de l'influence, sur la difficulté intrinsèque d'un item, de différents facteurs indépendants de la compétence évaluées. La performance à un item isolé ne permet pas de distinguer entre performance et compétence, donc de généraliser sur la maîtrise de l'objectif.

Nous illustrerons ce problème à partir d'un exemple simpliste, celui d'un item évaluant l'addition. Cet item peut être proposé sous la forme habituelle d'une addition, ou sous la forme « 6 + 5 = ? », ou sous la forme « Jean avait six francs, sa mère lui en a donné cinq de plus, combien possède-t-il en tout ? ». Bien que la tâche soit la même, la seconde présentation peut être déroutante pour des élèves habitués à la première. La troisième peut faire échouer des élèves maîtrisant l'addition mais ayant des difficultés de lecture ou de compréhension verbale. Que se passerait-il si on voulait évaluer la maîtrise de l'addition à partir d'un seul item ? On voit que dans le premier cas, on ne peut conclure sur la capacité des élèves à généraliser l'utilisation de la compétence à d'autres situations que la situation prototypique de l'opération telle qu'ils ont appris à la poser ; dans le second cas, on peut penser que le niveau de difficulté de l'item ne dépend pas que de la maîtrise de l'addition mais aussi de la capacité à transférer à une situation nouvelle ; et enfin, dans le troisième cas, la difficulté de l'item dépend d'une variable parasite, la compétence en lecture.

Pour pallier cet inconvénient, il est recommandé d'utiliser un nombre d'items « variés » suffisamment important pour neutraliser les biais liés aux contenus spécifiques d'un item. Un item isolé ne renseigne, à la limite, que sur lui-même ; les résultats à un ensemble d'items permettent d'inférer sur ce qui leur est commun : en principe la maîtrise d'un objectif. Nous retrouvons ici les idées de base de la psychométrie : un seul item ne permet pas d'apprécier l'erreur de mesure, plusieurs items évaluant la même compétence permettent de construire une mesure de la compétence du sujet (variable non observée).

Nous insistons sur ce point pour deux raisons. La première est que nous avons rencontré des situations où un objectif est évalué par très peu d'items, voire un seul. La seconde est que dans le cas où ces items sont de mauvaise qualité, comme nous l'avons parfois constaté au cours de nos analyses, l'évaluation d'un objectif à partir d'un ensemble d'items peu fiables, n'est pas valide. A fortiori, conduire des comparaisons temporelles entre items ou entre objectifs évalués par un faible nombre d'items, ne paraît pas offrir les garanties nécessaires pour construire des inférences valides.

1.4 - LA PUBLICATION DES RESULTATS

Nous notons des différences dans la publication des résultats entre les quatre dispositifs d'évaluations :

- en 1984, un volume par matière a été publié. On y trouve, dans l'ordre : des informations générales sur la construction du protocole, un rappel des programmes, la description des objectifs, puis des tableaux synthétiques présentant la réussite des items regroupés par objectifs et, enfin, les résultats à chaque item donnés de manière très complète, sur deux pages, accompagnés des informations recueillies auprès des enseignants (importance de l'objectif, pourcentage de réussite attendu). Nous notons qu'on ne calcule pas d'indicateurs synthétiques pour chacun des objectifs. A la fin de l'ouvrage sont présentés comme des annexes techniques les caractéristiques de l'échantillon et des points méthodologiques

(traitement des réponses des professeurs). Nous déplorons que les informations psychométriques telles que la consistance interne et la discrimination des items ne soient pas fournies.

- En 1990, aucune publication n'a été diffusée. Nous avons eu accès à des rapports internes. Il s'agit de rapports globaux portant sur l'ensemble des disciplines, un pour la troisième générale, un pour la troisième technologique. Les résultats sont présentés de manière synthétique par objectifs.
- Le rapport de 1995 rend compte en un volume de l'ensemble des disciplines. On trouve après une présentation générale, une présentation par discipline. Pour chaque discipline sont présentés les objectifs, les résultats par objectifs accompagnés de commentaires et illustrés par quelques items. En fin de volume, des « approfondissements » sont présentés, concernant les résultats d'études comparatives : temporelles, troisièmes générales et technologiques, LV1 et LV2, évaluation et brevet. Enfin, des annexes développent des points méthodologiques, en particulier les problèmes d'échantillonnage des sujets.
- L'évaluation de 1999 n'a pour l'instant donné lieu à aucune publication. Nous avons eu accès à des notes internes sur des points techniques (en particulier des études sur la fiabilité des comparaisons temporelles avec l'évaluation de 1995) et des rapports faits par des étudiants en stage à la DPD.

La formule par matière de 1984 nous semble motivante pour les utilisateurs fonctionnant dans un cadre disciplinaire. Cette formule permet de publier sous un volume maniable l'ensemble des informations sur le protocole d'une discipline. Les informations - à l'exception des indicateurs psychométriques - sont très complètes. On peut regretter l'absence de synthèse au niveau des objectifs. Le rapport 1995 offre, par contre, ce type d'approche. Cependant, la manière dont ces indicateurs sont calculés entraîne les défauts que nous avons signalés : nombre différent d'items par objectif, objectifs comportant un nombre trop faible d'items pour être évalués de façon fiable, absence de vérification de la consistance interne des groupes d'items rassemblés pour calculer un score par objectif. En conclusion, chacune de ces formules a des avantages et des inconvénients quant à la communication des résultats des évaluations-bilans en fin de collège. Si nous nous tournons vers des utilisateurs hors du système éducatif ou des utilisateurs fonctionnant hors du cadre des disciplines, les résultats analytiques par items et même par objectifs peuvent être trop techniques pour être aisément utilisables.

Il nous paraîtrait important de réfléchir à d'autres types de traitement et de présentation des résultats. Les documents publiés par différents organismes étrangers pourraient fournir dans ce domaine une base de réflexion intéressante. Il s'agit de considérer les objectifs et les items qui servent à les évaluer dans une perspective globale. Les items sont hiérarchisés selon leur difficulté. On peut donc les ordonner sur une échelle. On s'appuie alors sur cette propriété des protocoles pour présenter les items selon leurs niveaux de difficulté.

Nous avons constaté que de nombreux dispositifs (par exemple les enquêtes américaines National Assessment of Education in Progress, les enquêtes menées sur les acquis des élèves hollandais par CITO) utilisent les modèles de réponse à l'item (MRI) pour le traitement et la présentation des résultats. Les MRI sont des modèles puissants pour assurer la mesure en éducation (pour une présentation, voir Dickes et al., 1994 ; Vrignaud, 1996). Leur avantage est de placer conjointement sur une même échelle compétence des sujets et difficulté des items. La mesure de la compétence d'un sujet permet d'estimer dans quelle mesure il peut réussir un item d'une difficulté donnée. La mesure de la difficulté d'un item permet d'estimer le niveau de compétence requis pour pouvoir le réussir.

En fait, il serait plus exact de parler des effets de ces caractéristiques sur les performances des élèves. On cherche à estimer l'effet de différentes variables sur la performance des élèves. Ces analyses sont intéressantes car elles mettent bien en évidence l'effet de ces variables et attirent l'attention des éducateurs sur des faits auxquels il conviendrait, dans un souci d'équité, de remédier. Nous ferons quelques commentaires quant à la manière dont l'étude de ces effets est conduite.

1.5.1 - Constater n'est pas expliquer

En premier lieu, en ce qui concerne l'étude de l'effet de variables indépendantes de l'école, on est conduit à se demander s'il n'y a pas une forme de complaisance à répéter régulièrement des constats bien connus depuis les analyses sociologiques des années soixante. Il serait utile d'aller au-delà de ce constat en essayant de mieux comprendre en quoi ces variables agissent sur les performances scolaires afin d'y remédier plus directement. Nous pensons par exemple dans le domaine de la psychologie cognitive aux travaux de Lautrey sur l'influence des pratiques éducatives sur le développement cognitif (Lautrey, 1982) ; dans le domaine de la psychologie sociale, aux travaux de Monteil montrant l'influence du contexte social sur les performances cognitives (Monteil & Huguet, 1999) ; dans le domaine de la sociologie, aux travaux de Charlot (2001) et de son équipe, sur l'identité et le rapport au savoir.

D'autres études menées par la DPD vont effectivement dans ce sens : par exemple, l'étude des méthodes de travail ou l'étude de l'évolution socio-cognitive des élèves. On peut également ici s'interroger sur la variabilité inter et intra-individuelle. En effet, étudier un effet PCS conduit à ne pas s'interroger sur les différences existant entre individus appartenant à une même catégorie. Il serait pertinent de s'interroger sur la variabilité des parcours des sujets d'une même PCS comme le montrent par exemple les travaux du sociologue Bernard Lahire (1998). Il n'est bien sûr pas question de transformer la DPD en laboratoire de recherches, mais plutôt d'attirer l'attention sur la présentation un peu « mécanique » et statique, de ces mises en relation des caractéristiques des élèves avec leurs performances.

1.5.2 - Identifier les biais

L'étude des caractéristiques revient, de fait, à comparer les performances de groupes de sujets possédant une caractéristique commune (par exemple le genre). Les comparaisons ne sont fiables que si les instruments utilisés pour conduire cette comparaison ne sont pas biaisés au détriment de l'une des populations comparées.

Dans ce domaine également, on ne se contentera pas de postuler l'absence de biais, on la vérifiera en menant les analyses nécessaires. Par exemple, au moins pour les cohortes de 1984 et 1990, on sait qu'il existait en troisième générale une différence dans la composition de l'échantillon selon le genre, dans la mesure où davantage de garçons s'étaient orientés vers d'autres voies. Par conséquent, la comparaison selon le genre se trouvait ici sujette à un biais d'échantillonnage.

Nous insistons particulièrement sur un type de biais que nous avons observé à plusieurs reprises dans les protocoles : des biais d'items. Les biais d'items sont des nuisances pour la qualité de la mesure psychométrique. En effet, on peut montrer qu'un item biaisé mesure autre chose que la variable qu'il est censé mesurer et que cette variable parasite favorise - ou défavorise - un des

groupes étudiés. Si nous nous situons dans le cadre de l'utilisation de tests ou de questionnaires pour des comparaisons entre groupes, l'équivalence entre deux épreuves doit être démontrée, en premier lieu, par l'absence de biais au niveau des items. Ces études devraient être systématiquement effectuées et leurs résultats publiés.

1.6 - LES COMPARAISONS TEMPORELLES ET INTERNATIONALES

Comparaisons temporelles et comparaisons internationales apportent des éléments d'informations essentiels pour évaluer les acquis des élèves. Les commentaires sur l'évolution du niveau des générations sont de toutes les polémiques sur l'école. La comparaison du niveau de nos élèves avec ceux d'autres pays est l'un des moyens disponibles pour se faire une idée de l'efficacité de notre système éducatif (Bonora & Bacher, 1986). En effet, nous avons montré que du fait de la dépendance de la mesure par rapport au choix des objectifs et à la difficulté des items, il est virtuellement impossible d'avoir une évaluation absolue des acquis de nos élèves. Les comparaisons temporelles et internationales peuvent, par contre, nous fournir des informations sur le niveau relatif de nos élèves, et permettre d'étudier les effets de certains facteurs d'efficacité du système, surtout dans des systèmes assez fortement centralisés comme le système français, où l'effet de nombreux aspects ne peut être mis en évidence du fait de l'absence de variabilité.

1.6.1 - Les comparaisons temporelles

Etant donné le nombre de publications portant sur ce thème, nous pouvons dire que les comparaisons temporelles sont une préoccupation majeure de la DPD. De nombreuses études, très pertinentes et méthodologiquement intéressantes, ont été conduites. Ainsi, dans le document édité par C. Thélot (1992), établi à partir d'un rapport au Premier Ministre intitulé « Que savons-nous des connaissances des élèves ?», la majorité des études porte sur les comparaisons temporelles.

En ce qui concerne les évaluations de type bilan, l'importance des comparaisons temporelles est soulignée dès les consignes aux enseignants et élèves. Leur objectif, n'est pas toujours clairement défini : s'agit-il d'apprécier une évolution en général ou, comme on le souligne dans d'autres documents, d'apprécier l'effet de mesures pédagogiques ? Il est essentiel que la perspective choisie soit claire car les méthodes pertinentes pour répondre à la première question ne sont pas les mêmes que celles qui permettent de répondre à la seconde. Nous nous centrerons, ici, sur la première question : la mesure des évolutions de cohortes au cours du temps.

Au niveau de la troisième, des comparaisons entre les trois cohortes (1984, 1990, 1995) ont été effectuées. Nous avons eu l'impression que les conditions nécessaires à leur mise en œuvre n'avaient, cependant, pas été clairement perçues. En effet, on ne peut comparer que des exercices et des items strictement identiques (ces conditions, nécessaires, n'étant malgré tout, pas suffisantes…). La DPD a systématiquement repris dans les protocoles des exercices des protocoles antérieurs. La proportion d'items considérés comme "communs" est très variable d'un protocole à l'autre entre 1984, 1990, 1995 et 1999. Elle peut atteindre 50% dans certaines disciplines. Malheureusement, du fait de modifications dans les exercices appelés communs, et les conditions de passation, ces données ne permettent qu'imparfaitement d'effectuer des comparaisons fiables du fait de l'influence de différents biais.

- **biais d'administration** : La comparaison entre les consignes des différents protocoles

montre des différences dans les conditions de passation qui peuvent produire des biais difficiles à apprécier. La durée des épreuves varie quasiment du simple – le protocole de 1999, une seule séquence d'une heure – au double – les protocoles de 1984, 1990 et 1995 comprenaient deux séquences d'une heure. Dans certains cas les passations étaient « rythmées » par exercices (ou groupes d'exercices). La durée attribuée à chaque exercice ou groupe d'exercice a, parfois, varié entre les protocoles des différentes années. Dans certains cas, on est passé à une durée globale pour la séquence, que l'élève pouvait gérer à sa guise. Par ailleurs, il existe des différences dans la présentation des épreuves aux élèves, au moins dans la mise en page et le titre des cahiers. Par exemple, en 1984 et 1990, le cahier s'intitule « Evaluation dans les collèges » et en 1999, l'épreuve s'intitule « Etude comparative en classe de troisième ». Ces détails peuvent modifier les représentations du dispositif par les élèves et avoir un effet sur leur motivation.

- **biais d'échantillon** : La composition de la population des élèves de troisième s'est trouvée progressivement modifiée entre les années 1980 et 1999 du fait, dans un premier temps, de la disparition de l'orientation en fin de cinquième puis, dans un second temps, de la disparition des troisièmes technologiques. La proportion d'une génération scolarisée en troisième générale[4] est ainsi passée de 66% en 1984, à 76% en 1976, 86% en 1995 et enfin à 100% en 1999. Ce point est longuement discuté et traité dans la présentation des résultats de 1995, mais il est difficile d'apprécier dans quelle mesure les méthodes palliatives mises en œuvre permettent, effectivement, de corriger l'importance des biais liés aux changements dans la composition de la population.

- **biais d'items** : Les changements de contextes (programmes, environnement culturel, informations) peuvent engendrer des biais d'items. La mise en œuvre de techniques d'identification des biais d'items nous a conduit à conclure qu'au moins 20% des items communs pouvaient être considérés comme biaisés. Une fois éliminés les items non absolument identiques et les items biaisés, il ne reste souvent pas plus de 25% d'items communs qui ne sont plus vraiment représentatifs du protocole et ne couvrent que partiellement les objectifs. Lorsque nous avons entrepris d'appliquer les méthodologies préconisées pour les comparaisons aux protocoles de mathématiques des années 1984, 1990 et 1995 (Bonora et Vrignaud, 1997), nous avons constaté (1) que le nombre d'items communs était réduit ; (2) qu'une proportion importante de ces items étaient biaisés (Nous présentons une synthèse de ces travaux en annexe 9). Nous avions conclu que nous étions à la limite de la faisabilité pour mener des opérations de comparaison dans ces conditions.

Tous ces éléments, rendent d'emblée les comparaisons peu fiables. Au niveau de la troisième, des comparaisons entre les trois cohortes (1984, 1990, 1995) ont été effectuées directement sur les fréquences de réussite des items « communs » et des objectifs. Cette manière de procéder ne correspond pas aux méthodes statistiques appropriées pour l'étude des comparaisons temporelles souvent basées sur des modèles de mesure comme les Modèles de Réponse à l'Item (voir paragraphe 4) ou les analyses factorielles.

Nous pensons que la difficulté de conduire des comparaisons temporelles fiables est réduite du fait de trois facteurs : (1) des évolutions du contexte éducatif ; (2) des changements de personnels parmi les responsables de la DPD et les constructeurs de tests, (3) d'une méconnaissance de la psychométrie. Le premier point (changements de programmes, modification de la composition des populations scolarisées) échappe à la DPD, tout au plus peut-on essayer de mettre en place des mécanismes de rattrapage, mais dont l'efficacité est souvent

[4] Sans tenir compte des élèves scolarisés en SEGPA.

limitée. Le second point peut être considéré comme un aléa. Ceci étant, ces incidents montrent l'importance, pour de telles opérations, qui reposent sur des procédures qui ne sont pas toujours écrites, de la conservation de la mémoire et de la transmission des savoir-faire. Nous développerons davantage le troisième point dans la partie « propositions » car les méthodologies des comparaisons temporelles ont été l'objet de nombreux et importants développements dans le domaine de l'évaluation des systèmes éducatifs. Les méthodes, souvent très sophistiquées, ont été intégrées récemment à la DPD à la suite de travaux d'expertise (voir par exemple les études de faisabilité sur les comparaisons des performances des élèves de sixième et de troisième (Bonora et Vrignaud, 1997) ou de primaire (Grisay, 1997b), ainsi que les travaux de Flieller sur l'évolution des performances intellectuelles des élèves du primaire (Flieller et al.), 1994).

L'analyse des quatre dispositifs démontre l'impossibilité de conduire des comparaisons temporelles fiables entre les protocoles du fait de deux biais majeurs affectant la représentativité des sujets (modification dans la composition de la population) et la représentativité des items (effectif trop faible d'items pour l'ensemble des objectifs). Nous avons signalé, de plus, de nombreuses causes de biais potentiels dans les consignes et l'administration. Seules quelques comparaisons pourraient être effectuées en se limitant à des sous-domaines.

1.6.2 - Les comparaisons internationales

Les résultats des comparaisons internationales sont une source d'information incontournable. Depuis l'enquête pionnière de l'IEA (Association Internationale pour l'Evaluation du rendement scolaire), dans les années soixante, à laquelle la France prit une part active, notre pays a participé à de nombreuses enquêtes internationales sur les acquis des élèves. La responsabilité de ces enquêtes a été, au départ, assurée pour la France par divers organismes (INETOP, INRP, CIEP). Depuis les années 90, la DPD en est maître d'œuvre, pour la France. Au niveau de fin de collège, nous citerons la troisième enquête sur les connaissances en sciences (TIMSS, *Third International Mathematics and Science Survey*) dont les résultats ont été publiés, et une enquête sur la littératie (PISA, Programme International de l'OCDE pour le Suivi des Acquis des élèves) dont les résultats sont en cours de traitement. Outre l'intérêt de situer les acquis de nos élèves et le rendement de notre système éducatif comparé à celui d'autres pays, ces enquêtes permettent une réflexion enrichissante sur nos pratiques d'évaluation. La nécessité de rechercher un cadre commun à un ensemble de pays aux conceptions pédagogiques parfois assez hétérogènes conduit à réfléchir sur une conception des objectifs différente de celle développée à partir de la seule analyse de nos programmes. La DPD a, de plus, développé une réflexion sur les méthodologies employées dans ce type d'enquêtes. Cette réflexion menée en liaison avec différents laboratoires universitaires (Dickes & Flieller, 1999 ; Guérin-Pace & Blum, 1999 ; Vrignaud & Bonora, 1998 ; Vrignaud, 2001) et d'autres institutions, a permis à la DPD d'acquérir une expertise dans ce domaine. Dans ce cadre, la DPD a piloté une enquête européenne sur les compétences en lecture des élèves de 16 ans dans quatre pays européens (projet SOCRATES, cf. Bonnet et al.), 2001.

1.7 - LES EVALUATIONS-BILANS DANS D'AUTRES PAYS

Notre revue de la littérature et notre enquête par courrier électronique, nous ont montré que la pratique des évaluations de type bilan était variable et qu'elle n'était pas encore généralisée. Elle est souvent récente (moins de dix, voire de cinq ans dans de nombreux pays). Nous avons également constaté que notre niveau troisième (neuvième année des systèmes étrangers) n'était pas toujours le niveau privilégié pour les évaluations de type bilan à l'étranger. On retrouve néanmoins l'idée d'évaluer aux grandes articulations des systèmes éducatifs. L'exposé sur les dispositifs étrangers qui suit, ne porte pas uniquement sur le niveau équivalent à la troisième.

Un des premiers dispositifs d'évaluation de type bilan de grande ampleur est, certainement, l'enquête américaine NAEP (*National Assessment of Educational Progress*) dont la première phase eut lieu en 1969. Cette enquête reproduite périodiquement a pris en compte différentes disciplines et différents niveaux scolaires. Le choix du Ministère américain n'a pas été de créer un service scientifique de haut niveau, mais de faire financer, par *l'Office for Educational Research and Improvment*, des organismes scientifiques chargés de piloter l'ensemble de l'enquête jusqu'à la publication des résultats. Depuis 1983, cette mission est confiée à ETS (*Educational Testing Service*), le plus important organisme mondial dans le domaine de l'évaluation et de la recherche (signalons qu'ETS est un organisme privé). La NAEP applique par son ancienneté et les travaux scientifiques et méthodologiques qui ont accompagné son développement, comme un dispositif prototypique. La NAEP comporte une méthodologie basée sur les Modèles de Réponses aux Items (MRI) pour le traitement et la présentation des résultats. Un soin particulier a été apporté à l'organisation des plans d'enquête et questionnaires en vue de permettre un suivi des performances dans le temps ainsi qu'un suivi du début de la scolarité à l'âge adulte. De plus, tant les résultats de la NAEP que ses méthodologies ont fait l'objet de nombreuses publications. Par ailleurs, ETS a tenté d'en faire une référence pour les enquêtes internationales.

Parmi les organismes pionniers en matière d'évaluation des acquis des élèves, nous citerons également l'IEA, créée à la fin des années cinquante, qui lança sa première enquête internationale en 1961. L'IEA a joué un rôle essentiel dans l'élaboration des méthodologies d'évaluation des acquis des élèves et dans leur transmission aux différents organismes nationaux.

Sans chercher à être exhaustifs, nous citerons quelques exemples représentatifs des différentes situations rencontrées. En Hollande, CITO, un organisme privé, pratique, à la demande du Ministère de l'Education néerlandais, des évaluations de type bilan. Ces évaluations fonctionnent selon le modèle de la NAEP. La Finlande a mis en place, récemment, des évaluations qui, en plus des acquis scolaires, portent une attention particulière aux développement des apprentissages tout au long de la vie (apprendre à apprendre). Le cas de la Grande-Bretagne est particulier dans la mesure où les résultats individuels servent à évaluer tant l'élève (évaluation certificative) que les établissements. Quelques pays ne procèdent pas à des évaluations systématiques. Dans les pays fédéraux comme l'Allemagne ou la Suisse, la mise en œuvre et l'importance des dispositifs d'évaluations-bilans varient selon les initiatives régionales.

L'ancienneté du SIGES/SPRESE/DEP/DPD place la France parmi les pays ayant une longue tradition des évaluations de type bilan. Les dispositifs d'évaluation sont bien intégrés au système éducatif. Le service existe au sein du Ministère et fait appel à des organismes extérieurs pour des expertises ponctuelles. L'une de ses spécificités, par rapport à la plupart des organismes étrangers, est d'avoir mis en place des évaluations diagnostiques utilisables par les enseignants dans leur classe (culture de l'évaluation) ainsi que des recherches de remédiation.
Comparé aux dispositifs étrangers les plus sophistiqués, son principal handicap est sans doute l'absence de perspective psychométrique dans les évaluations de type bilan.

2 - PROPOSITIONS

2.1 - MIEUX IDENTIFIER ET DISTINGUER LES OBJECTIFS DES EVALUATIONS

La distinction des objectifs des évaluations est fondamentale car, d'une part les principes de construction, et par suite la qualité de chacune des épreuves, sont dépendants de leur finalité ;

car d'autre part l'utilisation d'épreuves à tout faire peut induire, lors de la communication des résultats aux utilisateurs, aux décideurs, au grand public, des représentations erronées sur ce que mesure chacun de ces types d'évaluation. Nous présentons une discussion approfondie de l'importance de cette distinction en annexe 6.

2.2 - AMELIORER LES EVALUATIONS DE TYPE BILAN

Les évaluations de type bilan sont l'instrument privilégié pour un observatoire des acquis des élèves. Leur objectif premier est de recueillir les informations les plus fiables possibles sur cette question des acquis. Ces informations doivent pouvoir être utilisées pour répondre à un ensemble de questions que l'on peut se poser sur les connaissances des élèves. Toutes les décisions lors de la construction des protocoles doivent être subordonnées à un objectif : assurer la mesure[5] !

2.2.1 - Les objets des évaluations : clarifier et étendre les domaines

2.2.1.1 - Evaluer les acquis de troisième, ou des quatre années de collège, ou de l'ensemble de la scolarité ?

Les épreuves de la DPD remplissent plutôt la première et la seconde fonction. Faut-il les faire évoluer vers les acquis scolaires en général ? Cette solution a l'avantage d'élargir la perspective, d'aller vers des compétences qui pourraient être mises en relation avec l'éducation et l'évaluation tout au long de la vie.

Cette extension conduirait, du moins pour les disciplines qui sont enseignées avant le collège, à définir les objectifs évalués de manière à ce qu'ils couvrent tout le parcours scolaire. Il serait utile de réfléchir à l'intérêt de situer les objectifs pris en compte par nos différentes évaluations nationales sur un continuum de la maternelle à la troisième, en passant par le primaire. Cette question est conceptuellement et méthodologiquement difficile.

2.2.1.2 - Evaluer toutes les disciplines ?

Nous avons vu que la plupart des disciplines (à l'exception de la musique et de l'éducation physique) ont fait l'objet d'au moins une évaluation. Cependant, certaines disciplines ont été évaluées plus fréquemment que d'autres. Si on envisage une périodicité quinquennale des évaluations de type bilan, alors deux ou trois disciplines pourraient être évaluées chaque année, ce qui éviterait une trop importante lourdeur des dispositifs.

2.2.1.3 - Evaluer quels apprentissages ?

La procédure classique de construction des items d'une épreuve consiste à élaborer un référentiel, extrapolé des textes officiels, qui croise le plus souvent les contenus disciplinaires et les démarches intellectuelles applicables sur ces contenus. Or, cette procédure pourra parfois apparaître insuffisante pour rendre compte exhaustivement des acquis de l'élève. Il semble que la conception d'un instrument d'évaluation puisse tirer profit d'une réflexion préalable sur les acquis les plus significatifs de la formation considérée, en particulier ceux qui transcendent la visée disciplinaire.

Outre une recherche des objectifs inter ou transdisciplinaires, ou portant davantage sur les

[5] Selon le titre de l'ouvrage de Cardinet et Tourneur, 1980.

savoir-faire et moins sur les contenus, qui ont déjà fait l'objet de séries d'items intégrés dans certains protocoles d'évaluation du MEN, on peut penser en particulier à l'intérêt que pourrait présenter l'évaluation des acquis les plus **transférables**, soit dans le cadre de la discipline elle-même (facilitation des acquis ultérieurs), soit dans un cadre plus large (vie quotidienne, professionnelle, etc.).

On pourrait aussi envisager les acquis en termes de structuration des connaissances, et tenter d'évaluer le niveau atteint par les sujets dans la structuration et l'assimilation des notions-clés de la discipline, notions qui peuvent être saturées de contenus (e.g. notion d'énergie...), ou impliquer plus directement des savoir-faire (e.g. notion de réversibilité...).

Il serait également intéressant de dégager des pistes d'élaboration des items à partir de la prise en compte des "critères d'intelligibilité" de la discipline, qui évoluent dans le temps, et modifient les caractéristiques générales des compétences de la population (e.g. l'enseignement du français, vu plutôt comme étude de la littérature, puis comme amélioration de la compréhension de la langue, puis comme étude des spécificités des différentes formes d'écrit (cf. Bonora, 1996 ; Develay, 1996) ; de même l'enseignement des sciences naturelles, vu plutôt comme étude des structures anatomiques et des classifications, puis comme étude des fonctionnements biologiques...). Les compétences des sujets étant dépendantes de ces changements, la prise en considération de telles évolutions des "matrices disciplinaires" devrait conduire aussi à pondérer les résultats des comparaisons temporelles des niveaux atteints par les populations.

2.2.1.4 - Evaluer les productions écrites ?

Des productions d'écrits (voire des productions de dessins dans l'épreuve d'arts plastiques) ont été demandées dans les protocoles de la DPD. Ce point est relativement original par rapport aux protocoles d'évaluation de nombreux organismes qui se limitent à des QCM ou à des questions ouvertes avec des réponses brèves. Ces exercices sont également utiles car ils renforcent la validité de contenu des protocoles. Le travail de cotation de telles productions est bien sûr plus lourd que celui des QCM. Il doit faire l'objet d'une élaboration de grilles de dépouillement à partir d'indicateurs factuels à défaut de quoi il n'échapperait pas aux biais habituels liés à la subjectivité des correcteurs. Nous avons constaté que ces grilles se sont progressivement réduites au fil du temps. Ces points mériteraient d'être approfondis. Signalons qu'actuellement des recherches sur la correction automatisée de textes produits, simulant la démarche d'enseignants, sont conduites en particulier dans des organismes tels qu'ETS aux Etats-Unis.

2.2.1.5 - Evaluer les productions orales ?

L'absence des productions orales dans la plupart des protocoles est souvent regrettée. Elle apparaît d'autant plus dommageable que les programmes mettent l'accent sur l'oral. De plus, les compétences de communication orale sont fondamentales dans la vie professionnelle, que ce soit en langue nationale ou en langues vivantes étrangères. Une évaluation standardisée de l'oral requiert des dispositifs encore plus complexes que l'évaluation des productions écrites. En effet, ces épreuves sont encore plus sensibles aux contextes psychologiques et psycho-sociaux. Il est évident que la connaissance préalable - en bien ou en mal - d'un élève ne peut qu'avoir un effet parasite sur l'appréciation de sa prestation par l'enseignant. Les situations doivent non seulement être standardisées mais aussi ne pas induire de prises de position trop marquées chez l'élève. Ce type d'instruments mériterait également des études de fiabilité avant qu'ils ne soient introduits et systématisés dans les protocoles.

2.2.1.6 - Evaluer les connaissances transdisciplinaires ?

Les connaissances disciplinaires ne sont qu'un aspect de l'enseignement de l'école et du collège. Méthode de travail, « apprendre à apprendre », font aussi partie des objectifs, enseignés de manière « latente ». L'évaluation de ces savoir-faire a donné lieu à plusieurs instruments intégrés dans des protocoles d'évaluation : un protocole transdisciplinaire dans l'évaluation de 1990 et surtout les questionnaires sur les méthodes de travail. Nous rattacherons également, à ce pôle transdisciplinaire, l'évaluation des compétences en informatique.

Il est important d'évaluer ces connaissances car elles sont transférables à d'autres situations de formation, et à long terme dans la vie professionnelle. Ces savoirs sont d'autant plus intéressants qu'ils ont souvent un pouvoir explicatif important sur l'ensemble des performances disciplinaires et qu'il est possible d'agir sur leur acquisition.

2.2.1.7 - Evaluer des compétences « professionnelles » ?

Une définition des objectifs en termes de compétences professionnelles pourrait être intéressante, quoique d'aucuns jugeront cette approche prématurée dans la mesure où la plupart des collégiens poursuivront leur formation durant encore plusieurs années. Le terme de compétences utilisé actuellement dans les deux univers (DPD et emploi) ne doit pas faire illusion : il est impossible d'établir une relation entre les objectifs, les compétences pédagogiques des évaluations actuelles et des éléments se rapprochant de ce qui est évalué dans le monde professionnel sous cette dernière appellation. Depuis plus d'une vingtaine d'années, la définition des compétences, leur identification chez les personnes, voire la description des emplois en termes de compétences, a fait couler beaucoup d'encre. Sans envisager des programmes lourds, il serait cependant utile de tenter d'établir un système permettant de mettre en relation les objectifs pédagogiques évalués et les compétences professionnelles.

2.2.1.8 - Evaluer les compétences sociales ?

Nous ne définirons pas les compétences sociales, cette question pourrait, à elle seule, faire l'objet d'un rapport. De nombreux travaux conduits dans le cadre des dispositifs de bilan, en particulier du bilan des jeunes, ont cherché à les définir et à les identifier. L'évaluation standardisée de ces compétences est délicate. Certains contenus du questionnaire sur la vie scolaire, auquel nous faisions allusion plus haut, comportent des dimensions qui peuvent tout à fait être assimilées à ce domaine. Ce point nécessiterait des recherches spécifiques.

2.2.2 - Les protocoles d'évaluation : renforcer la fiabilité de la mesure

2.2.2.1 - Augmenter le nombre d'items par objectif

Cette recommandation traverse notre rapport comme un *leitmotiv* car elle est fondamentale et nous est apparue bien imparfaitement respectée. Répétons le une fois encore, seule l'utilisation de plusieurs items ou plutôt exercices peut permettre de généraliser sur un objectif donné.

Cette exigence peut apparaître contraignante par rapport à l'étendue des objectifs à évaluer et au temps de passation disponible. La solution adoptée dans plusieurs protocoles de 1984 et 1990, et par de nombreux organismes étrangers, consiste à mettre en place un système de blocs d'items ou plutôt d'exercices. Chaque sujet ne passe qu'une partie des blocs. Un dispositif d'ancrage permet de placer tous les blocs sur une même échelle, soit à partir d'un bloc commun, soit en utilisant une combinatoire permettant d'établir des relations entre tous les blocs deux à deux

(système dit des plans expérimentaux incomplets). Ces dispositifs permettent de recueillir des informations sur un grand nombre d'items sans augmenter trop la charge de travail des sujets.

2.2.2.2 - Réduire les sources de biais

- Minimiser les liaisons artificielles entre items

  Les protocoles de la DPD comportent des exercices comprenant plusieurs items. A l'intérieur d'un même exercice, les réponses des élèves ne sont, le plus souvent, pas indépendantes les unes des autres. Ce fonctionnement induit une cohérence artificielle dont il faudrait tenir compte dans l'appréciation des indicateurs psychométriques.

  Il serait judicieux d'éviter ce type de situations débouchant sur des items dont la forme crée une dépendance locale forte : par exemple, les réponses en « cascade » où le sujet doit, dans un premier temps, fournir une réponse et, dans un second temps, la justifier, ainsi que les questions sous forme de classifications qui demandent au sujet de mettre en relations plusieurs termes avec des cibles.

- Clarifier le codage

  La distinction entre bonne et quasi-bonne réponse (codes 1 et 2) est sans doute intéressante pédagogiquement mais crée un flou certain du fait d'un écart parfois important dans les fréquences de réussite selon que l'on inclut les codes 2 ou non. Or, rappelons le, les items sont censés informer sur l'acquisition de l'objectif, il faut donc décider clairement si les quasi-bonnes réponses indiquent ou non cette maîtrise.

2.2.2.3 - Vérifier la qualité du protocole par une réelle pré-expérimentation

Une pré-expérimentation est la seule procédure fiable pour vérifier les qualités de l'épreuve. Ce n'est pas une simple formalité, les résultats des analyses conduiront à supprimer, éventuellement à modifier, les exercices, les items et leur codage. L'étude des indicateurs de difficulté permettra de fixer le niveau de difficulté global souhaité pour l'épreuve. Rappelons que les jugements d'experts ne sont en aucun cas fiables, comme l'ont montré toutes les comparaisons antérieures effectuées entre jugements et difficulté observée. De même, discrimination et consistance sont les seuls indicateurs objectifs du rattachement ou non des exercices et des items aux objectifs. Enfin, on pourra effectuer certaines vérifications telles que l'identification des items biaisés selon les caractéristiques des élèves.

Dans les cas où l'on souhaite employer des modèles particuliers pour le traitement des résultats (par exemple les MRI), il sera nécessaire de s'intéresser lors de l'analyse des résultats de la pré-expérimentation, à l'adéquation du modèle aux données, et d'éliminer, si nécessaire, les items témoignant d'une insuffisante adéquation.

Cette pré-expérimentation ne dispensera pas, bien entendu, de conduire les mêmes analyses sur le protocole définitif.

2.2.2.4. Mieux standardiser les conditions de passation

Nous avons constaté, en particulier, dans le déroulement des passations (gestion du temps, lecture des consignes), une variabilité entre les classes, et entre les établissements. Cette variabilité devrait être éliminée autant que faire se peut.

En ce qui concerne les évaluations de type bilan, nous sommes favorables à une solution qui éviterait l'administration des épreuves par les professeurs habituels, ceci pour au moins deux raisons. La première est qu'une telle solution permettrait de réduire d'éventuels effets induits par la relation développée au cours de l'année scolaire, et de mieux respecter l'anonymat signalé dans les consignes lues aux élèves. La seconde est qu'elle limite l'éventualité d'une divulgation de tout ou partie des épreuves, point sur lequel nous reviendrons à propos de l'étude des évolutions temporelles. Certains organismes étrangers font administrer les protocoles des évaluations de type bilan par des examinateurs spécialisés.

2.2.2.5 - Augmenter la taille des échantillons

La taille des échantillons est suffisante en ce qui concerne les traitements statistiques actuellement mis en œuvre. Il serait, par contre, nécessaire de les augmenter dans le cas où on déciderait d'utiliser d'autres traitements, tels que les modèles de la réponse à l'item, pour lesquels l'estimation des paramètres nécessite une taille plus importante des échantillons.

2.3 - LA COMMUNICATION DES RESULTATS : SYSTEMATISER ET RENOUVELER LA PRESENTATION

2.3.1 - Les exercices

La publication du texte intégral des exercices n'a été faite que pour le dispositif de 1984. La publication de l'intégralité des exercices n'est pas forcément indispensable, un choix d'exercices typiques d'un objectif ou d'un niveau peut être suffisant. Signalons que, dans un souci de réduction des biais, de nombreux organismes ne divulguent pas l'ensemble des exercices susceptibles d'être utilisés pour des comparaisons temporelles.

2.3.2. - Les indicateurs psychométriques

Nous insistons fortement sur les raisons scientifiques et déontologiques de publier l'intégralité des indicateurs psychométriques (difficulté, discrimination, consistance) pour les items et l'ensemble du protocole. Cette publication est indispensable pour permettre un regard scientifique sur la fiabilité de la mesure. Elle fait partie des procédures recommandées par les organismes garants de la qualité des tests et instruments d'évaluation (par exemple la Commission Internationale des Tests). Ces indicateurs sont bien sûr accompagnés des commentaires nécessaires, sans oublier de signaler les choix qui ont pu être faits, leurs raisons et leurs incidences : suppression d'un ou de plusieurs items, fusion de codages, maintien d'items de qualité médiocre, etc. Ces informations peuvent, bien sûr, ne faire l'objet que d'un rapport technique diffusé aux seuls spécialistes. Il est, évidemment, nécessaire de signaler dans les documents pour un public plus large, les biais éventuels limitant la généralisation des résultats des items et des exercices problématiques.

2.3.3 - Les résultats

La DPD produit plusieurs publications (Dossiers d'Education et Formation, Notes d'Information), dont la longueur, le niveau de détail et la largeur de la diffusion offrent une diversité suffisante, ce que nous considérons comme très satisfaisant.

Le principal problème nous semble concerner les types d'indicateurs publiés, en particulier la fréquence de réussite des items. La présentation sous forme de fréquences a le mérite de la simplicité. Elle a l'inconvénient d'induire une réification et surtout d'inciter à des comparaisons qui ne sont pas toujours fiables. En effet, nous avons rappelé que le problème principal de la mesure en éducation est la difficulté de distinguer la réussite de la tâche et la compétence des sujets.

D'autres modes de présentation pourraient être envisagés. Le recours aux modèles de réponse à l'item a été souvent adopté par des organismes étrangers et internationaux d'évaluation en éducation. L'avantage est de placer la difficulté des items et la compétence des sujets sur une même échelle. On peut donc lire sur un même graphique la difficulté des items et la fréquence des sujets ayant une compétence correspondant à cette maîtrise. De plus on peut indiquer les niveaux de compétence de différents groupes (selon le genre, la PCS, etc.). Signalons enfin, que ces modèles permettent, à condition d'avoir mis en place un système d'ancrage approprié, d'établir des comparaisons entre les niveaux des sujets des différentes cohortes. On voit l'avantage et l'élégance de cette présentation synthétique. L'inconvénient de ces modèles tient aux contraintes qu'ils imposent (exigences d'unidimensionnalité, d'indépendance locale, etc.). Une réflexion devrait être conduite sur ce point. Il serait également intéressant d'envisager l'utilisation d'autres modèles de mesure prenant en compte la multidimensionnalité, comme par exemple l'analyse géométrique des données.

2.3.4 - Un contrôle scientifique des publications ?

Pour terminer, il nous semblerait souhaitable que soit entreprise une réflexion sur le contrôle de la qualité scientifique des publications de la DPD. Cette réflexion est motivée par la position particulière de la DPD : service d'une administration élaborant un produit scientifique. Dans les publications scientifiques, le contrôle par les pairs est de règle. Les articles dans des revues à comité de lecture sont révisés par plusieurs experts qui peuvent autoriser ou non la publication et surtout demander des modifications et/ou des précisions. Nous avons rencontré dans les publications de la DPD (revue « Education et Formation ») des articles qui auraient pu être nettement améliorés par des expertises extérieures (par exemple le manque de références aux travaux antérieurs sur le thème traité par l'article).

2.4 - LES COMPARAISONS TEMPORELLES : MIEUX GERER LE DISPOSITIF

L'évolution temporelle est une préoccupation majeure de la DPD. Un observatoire du système éducatif doit nécessairement être en mesure d'informer sur les changements dans les acquis de nos élèves au cours du temps. Nous pensons que l'importance de l'enjeu mérite d'y consacrer des moyens importants.

Cet objectif étant prioritaire, il est indispensable de mettre en place toutes les mesures pour assurer la fiabilité de ces comparaisons. Satisfaire cet objectif introduira des contraintes fortes. Il existe une abondante littérature psychométrique sur les procédures et les méthodes à mettre en œuvre pour construire et gérer un dispositif permettant la comparaisons des acquis des élèves dans le temps. Leur mise en place est une entreprise exigeante demandant de respecter de nombreuses contraintes au niveau du dispositif de production des données, en particulier les protocoles. Il est évidemment indispensable d'utiliser les mêmes exercices, les mêmes items sans les modifier, même d'une virgule, et, sans modifier les conditions de passation (consignes, durée de travail, place de l'exercice dans le protocole).

Comparer dans le temps, veut dire pouvoir placer sur une même échelle de mesure, des résultats obtenus par des générations différentes, dans un contexte sujet aux changements. Ces changements impliquent qu'il est rarement possible d'administrer des protocoles identiques car certains exercices « vieillissent » plus mal que d'autres. La solution la plus largement préconisée est d'utiliser une partie des exercices pour assurer ce qu'on appelle en termes techniques un ancrage. On peut ainsi ne conserver dans les protocoles qu'une partie des exercices acceptables dans le nouveau contexte.

Construire un tel dispositif rend nécessaire de disposer d'un ensemble important d'exercices et d'items pour assurer cet ancrage. Précisons que ces items d'ancrage ne doivent pas forcément être utilisés indéfiniment : il suffit d'assurer un ancrage entre deux évaluations consécutives. Ces exercices doivent évidemment être représentatifs des contenus et aussi des niveaux de difficulté, et bien sûr, être en nombre suffisant.

Il sera également utile d'envisager des formations spécifiques à ces méthodes. Nous avons constaté que les statisticiens de la DPD, issus de l'ENSAE, n'étaient pas formés à ces méthodes (signalons que les personnels actuellement en place à la DPD ont suivi des formations spécifiques dans le cadre de la formation continue).

Ces problèmes techniques une fois résolus, il reste qu'au niveau des interprétations, un travail important d'élucidation, qui devra intégrer les évolutions entre les moments successifs des évaluations, sera nécessaire : évolution dans les programmes scolaires, mais aussi dans les environnements social et technique, qui modifient l'information, ainsi que le mode d'accès à cette information, et les attitudes.

2.5 - LE RECOURS A L'INFORMATIQUE, DE NOUVELLES POSSIBILITES

2.5.1 - Construire et gérer des banques d'item

La construction d'épreuves d'évaluation est largement facilitée aujourd'hui par la construction et la gestion de banques d'items. On définit la banque d'item comme « *une collection d'items organisés, classés et catalogués, tels des livres dans une librairie, en vue de faciliter la construction d'une grande variété de tests de performance et d'autres types de tests mentaux* » (Choppin, 1988, cité par Dickes et al, 1994, page 76).

La banque d'items sert à construire des tests sur mesure (sur ce thème, voir Vrignaud, 1996, 2000 ; Vrignaud & Chartier, 1999). On emploie le terme "sur mesure" (qui rend l'anglais *tailored*) pour désigner des tests dont les items sont choisis pour approcher au mieux, quant à leur difficulté, le niveau de compétence de la population évaluée, estimé grossièrement dans une phase initiale (on se rapproche là du test adaptatif, sur lequel nous reviendrons). Les spécifications du constructeur sont respectées en retenant dans la banque les items les plus appropriés selon les valeurs de leurs paramètres psychométriques (e.g. difficulté, discrimination) qui ont été au préalable estimés (calibration) sur un échantillon, puis entrés dans la banque d'items. Les banques d'items répondent bien aux problèmes suscités par le choix des items d'ancrage pour l'étude des évolutions temporelles, ou pour s'assurer que le niveau d'un protocole est comparable d'une session à l'autre. Par exemple, ETS utilise une banque d'items pour élaborer chaque année de nouvelles versions du TOEFL (*Test Of English as a Foreign Language*) en garantissant la comparabilité du niveau attribué aux candidats d'une année à l'autre. En France, le CIEP (Centre International d'Etudes Pédagogiques) a élaboré, sur ce principe, une banque d'items pour construire des tests de français comme langue étrangère.

La construction d'une banque d'items est un travail lourd et de longue haleine. Les organismes qui utilisent ces méthodes disposent souvent de plusieurs milliers d'items calibrés, de bonne qualité, pour une discipline donnée. Des logiciels particuliers sont dédiés à la construction et à la gestion des banques d'items. Les banques d'items représentent, selon nous, la solution la plus adaptée au suivi dans le temps et à l'élaboration d'épreuves d'ancrage pour le brevet.
Actuellement, la DPD a publié une banque d'exercices sur support papier qui sera progressivement mise à disposition sur un site web. Il ne s'agit donc pas d'une banque d'items au sens habituel du terme puisque l'objectif n'est pas de construire des protocoles mais de disposer d'exercices d'évaluation dans la classe.

Par ailleurs, ces outils peuvent avoir une utilité pour l'harmonisation des épreuves de contrôle continu. Les enseignants peuvent consulter ainsi des exercices-types pour lesquels sont souvent disponibles des références (pourcentage de réussite à telle ou telle évaluation). Nous avons vu un système un peu similaire dans le Land de Nord Rhénanie - Westphalie qui diffuse auprès des enseignants des fascicules d'exercices-types.

2.5.2 - Mettre les protocoles sur support informatique

L'utilisation de l'informatique pour l'évaluation a suivi les évolutions technologiques : d'abord auxiliaire pour la saisie et le traitement des données, l'ordinateur est aujourd'hui un auxiliaire dans les passations. De nombreux tests et questionnaires sont présentés au sujet sur écran, les réponses saisies directement par le clavier et/ou la souris. Le format informatique ouvre de nouvelles possibilités, d'une part pour le format des items, d'autre part pour l'administration du test.

Au niveau du format, l'ordinateur en tant que support multimédia offre la possibilité d'utiliser images, son et vidéo. L'utilisation de ces supports peut être particulièrement attractive pour des disciplines (langues, disciplines artistiques, etc.) où la présentation de documents visuels ou sonores est indispensable pour construire les situations d'évaluation.

L'interactivité offerte par l'ordinateur permet de construire des tests adaptatifs, c'est à dire des tests dont les questions évoluent en fonction des réponses des sujets : si le sujet réussit l'item, il se voit proposer une question plus difficile, une plus facile, s'il échoue. Cette procédure présente deux avantages : (1) aboutir à une évaluation de la compétence du sujet avec moins d'items, (2) ne pas confronter le sujet à trop de situations où il échoue.

Ce champ d'exploration pourrait, à plus ou moins long terme, déboucher sur l'implantation de telles épreuves sur un site web utilisable par les enseignants et, pourquoi pas, par les élèves eux-mêmes. On pourrait, ainsi offrir une possibilité d'auto-évaluation à partir de ces protocoles sur support informatique, en particulier ceux du type des tests adaptatifs. Il faudrait bien sûr cadrer soigneusement ce dispositif et envisager la manière de communiquer les résultats à l'élève. Signalons que des recherches sont entreprises dans ce domaine au Luxembourg par l'ISERP.

3 - SYNTHESE

Tout d'abord, en dépit des critiques formulées, il importe de saluer l'intérêt et l'importance de l'outil que sont les évaluations de type bilan et les travaux de la DPD en général. Dans un ouvrage récent sur le « Le pilotage des systèmes éducatifs », Gilbert de Landsheere (1994) présente le dispositif français comme « un développement exemplaire ».

Nous pouvons dire que les dispositifs des évaluations de type bilan ont fourni et fournissent beaucoup d'informations sur les acquis des élèves en fin de collège en termes d'objectifs pédagogiques. En cela, ils ont rempli et remplissent bien leur fonction dans le cadre du pilotage du système éducatif. Ce centrage sur les objectifs pédagogiques devient une limite dans une perspective plus large, que ce soit celle de l'évaluation des acquis tout au long de la **scolarité** ou, encore plus, celle de l'évaluation tout au long de la **vie**. De même, les résultats publiés sont, selon nous, peu informatifs pour un public de non-spécialistes.

Nous avons longuement développé l'analyse de la fiabilité du dispositif. Nous avons montré que celui-ci était parfois insuffisant du point de vue de la qualité de la mesure. Nous avons pointé ce manque, en particulier, dans la mise en œuvre des comparaisons temporelles.

Nous avons présenté des propositions pour améliorer le dispositif : d'une part en prenant davantage de distance avec les seuls objectifs pédagogiques de la classe de troisième, d'autre part en améliorant la qualité de la mesure. Certaines propositions sont lourdes à mettre en œuvre : toutes celles qui concernent l'élargissement et le changement de définition des compétences évaluées (par exemple l'intégration des productions orales). D'autres, nous paraissent plus accessibles et nous les jugeons, d'ailleurs, plus urgentes : celles qui concernent l'amélioration de la fiabilité de la mesure et la mise en place d'un système solide permettant l'observation des évolutions temporelles.

REFERENCES

Baudelot, C. & Establet, R. (1991). Filles et garçons devant l'évaluation . *Education & Formations* ; *27-28* ; 49-66 .

Bonnet, G., & al. (2001). *The use of national reading tests for international comparisons: ways of overcoming cultural bias*. Socrates contract n° 98-01-3PE-0414-00.

Bonora, D., & Bacher, F. (1986). Intérêt et difficultés des comparaisons dans les enquêtes internationales, l'exemple de l'I.E.A*., Revue Tunisienne des Sciences de l'Education*, 11, n° 14, 129-138.

Bonora, D. (1996). Les modalités de l'évaluation. *Revue Internationale d'Education. Sévres*, *N° 11*, 69-85.

Bonora, D., & Vrignaud, P. (1997). *Evolution des connaissances scolaires et Modèles de Réponse à l'Item*. Rapport pour le Ministère de l'Education Nationale. Direction de l'Evaluation et de la Prospective.

Bourdon, J. & Thélot, C. (Eds) (2001) . *Education et formation : l'apport de la recherche aux politiques éducatives.* Paris : Editions du CNRS .

Cardinet, J., & Tourneur, (1980). *Assurer la mesure* . Berne : Peter Lang.

Charlot, B. (2001). Le rapport au savoir . in J. Bourdon & C. Thélot (Eds) . *Education et formation : l'apport de la recherche aux politiques éducatives* . pp. 17-34 Paris : Editions du CNRS .

Chartier, A. M. (1998). Epreuves du certificat d'études primaires en 1995. Etude de quelques facteurs ayant pu agir sur les résultats des élèves . *Education & Formations, 53* ; 19-34 .

Chartier, P., Vrignaud, P. (1999). *Analyse critique des banques d'outils d'aide à l'évaluation publiées par la DPD*. Rapport pour le Ministère de l'Education Nationale. Direction de la Programmation et du Développement.

de Landsheere, G. (1994). *Le pilotage des systèmes d'éducation*. Bruxelles : De Boeck.

Develay, M. (1996). Didactique et pédagogie. *Sciences Humaines*, *N°12 Hors Série « Eduquer et Former* », 58-60.

Dickes, P., Tournois, J., Flieller, A., & Kop, J.-L. (1994). *Psychométrie*, Paris: PUF.

Dubet, F., & Duru-Bellat, M. (2000). *L'hypocrisie scolaire. Pour un collège enfin démocratique.* Paris : Seuil.

Fayol et al. (2000). *maîtriser la lecture*. Publication de l'Observatoire National de la Lecture. Paris : Odile Jacob.

Flieller, A. (2001). Les compétences et les performances cognitives dans l'évaluation scolaire . in J. Bourdon & C. Thélot (Eds) . *Education et formation : l'apport de la recherche aux politiques éducatives* . pp. 187-200 Paris : Editions du CNRS .

Flieller, A., Manciaux, M., & Kop, J.-L. (1994). Evolution des compétences cognitives des élèves en début de scolarité élémentaire sur une période de vingt ans. *Les dossiers d'Education & Formations*, *47*, 205-218.

Guérin-Pace, (France), Blum, (Alain). L'illusion comparative. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme. *Population, 54,* 1999, p. 271-302.

Jarousse, J.-P., & Leroy-Audoin, C. (2001). Les nouveaux outils d'évaluation : quel intérêt pour l'analyse des "effets-classe" ? . in J. Bourdon & C. Thélot (Eds) . *Education et formation : l'apport de la recherche aux politiques éducatives* . pp. 163-185 Paris : Editions du CNRS .

Joutard, P., & Thélot, C. (1999). *Réussir l'école. Pour une politique éducative*. Paris : Seuil.

Grisay, A. (1997a). Evolution des acquis cognitifs et socio-affectifs des élèves au cours des années de collège. *Les dossiers d'Education & Formations*, *88*, août 1997.

Grisay, A. (1997b). *Etude par le modèle de Rasch de l'évolution des compétences des élèves de CE2*. Rapport pour le Ministère de l'Education Nationale. Direction de l'Evaluation et de la Prospective.

Lautrey, J. (1982). *Intelligence, hérédité, milieu*. Paris : PUF.

Lahire, B. (1998). *L'homme pluriel. Les ressorts de l'action*. Paris : Nathan.

Le Guen, M. (1991). L'évaluation CE2-6ème : un outil de connaissance des acquis des élèves . *Education & Formations* ; *27-28* ; 5-8 .

Levasseur, J., & Trussy, C. (1997). Parler en allemand, en anglais en français. *Les dossiers d'Education & Formations*, *91*, Septembre 1997.

Lieury, A. (1996). Mémoire encyclopédique & devenir scolaire : Etude longitudinale d'une cohorte sur les quatre années du collège français . *Psychologie et Psychométrie* ; *17* ; 33-44 .

Linn, R.L. (Ed) (1989). *Educational measurement (3rd ed.)* . New-York: Macmillan .

Monteil, J.-M. & Huguet, P. (1999). *Social context and cognitive performance : toward a social psychology of cognition.* London : Psychology Press.

Murat, F. (1998). Les différentes façons d'évaluer le niveau des élèves en fin de collège . *Education & Formations* ; *53* ; 35-50 .

Piéron, H. (1963). *La docimologie*. Paris : PUF.

Reuchlin, M., & Bacher, F. (1969). *L'orientation à la fin du premier cycle secondaire*. Paris : PUF.

Thélot, C. (Ed.). (1992) Que sait-on des connaissances des élèves ? *Les dossiers d'Education & Formations*, *17*, octobre 1992.

Tomassone, R. (1991). Diagnostic des élèves, diagnostic de l'évaluation . *Education & Formations* ; *27-28* ; 97-106 .

Vrignaud, P. (1996). Les tests au XXIéme siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes. *Pratiques Psychologiques, 2*, 5-28.

Vrignaud, P. (2000). Les banques d'items : une nouvelle réponse aux besoins d'évaluation ? *Biennale de l'Education*. Paris : 12-15 avril 2000.

Vrignaud, P. (2001). Evaluation sans frontières : comparaisons interculturelles et évaluations dans le domaine de la cognition. in M. Huteau. *Les figures de l'intelligence*. pp. 79-115 Paris : Editions et Applications Psychologiques .

Vrignaud, P., & Bonora, D. (1998). Literacy assessment and international comparisons. in Dan Wagner. *Literacy assessment for out-of-school youth and adults*. Philadelphia : UNESCO/International Literacy Institute.

Vrignaud, P., Chartier, P. (1999). *Phénix : Elaboration d'une méthodologie de construction et de gestion d'une banque d'items pour l'élaboration de tests sur mesure*. Paris : INETOP.

Vrignaud, P., Chartier, P., Dosnon, O. & Bonora, D. (1998). *Définition et mise au point d'indicateurs en vue de caractériser les items d'une banque d'outils d'aide à l'évaluation*. Rapport pour le Ministère de l'Education Nationale. Direction de la Programmation et du Développement.

Protocoles d'évaluation :

Evaluation 1984 :

Titre :
Evaluation dans les Collèges 1984 - fin du cycle d'orientation.
Service :
Ministère de l'Education Nationale. Service de la Prévision des Statistiques et de l'Evaluation. Direction des lycées et collèges. Division de l'Evaluation du Système Educatif. Département de l'évaluation dans les collèges et les lycées.
Disciplines évaluées :
Français, Mathématiques, Allemand (LV1), Anglais (LV1), Sciences Naturelles, Sciences Physiques, Histoire - Géographie.
Matériel :
Cahiers de l'élève et feuilles de réponse mécanographique ; cassette pour les LV(allemand, anglais) ; questionnaire « Vie Scolaire » ; Livret de consignes (application, codage) ; questionnaire enseignant ; relevé des réponses des élèves de la division.
Résultats :
Evaluation pédagogique dans les collèges. Fin de cycle d'orientation juin 1984. Ministère de l'Education Nationale. Service de la Prévision des Statistiques et de l'Evaluation. Direction des lycées et collèges. (une brochure par discipline).

Evaluation 1988 :

Titre :
Evaluation en fin de troisième technologique.

Service :
Ministère de l'Education Nationale. Service de la Prévision des Statistiques et de l'Evaluation. Direction des lycées et collèges. Division de l'Evaluation du Système Educatif. Département de l'évaluation dans les collèges et les lycées.
Disciplines évaluées :
Français, Mathématiques, Anglais (LV1), Epreuve transdisciplinaire.
Matériel :
Cahiers de l'élève et feuilles de réponse mécanographique ; cassette pour les LV(allemand, anglais) ; questionnaire « Vie Scolaire » ; Livret de consignes (application, codage) ; questionnaire enseignant ; relevé des réponses des élèves de la division.
Résultats :

(1989) Evaluation en fin de troisième technologique. Les dossiers d'Education & Formations, mai 1989.

Evaluation 1990 troisième Générale:

Titre :
Evaluation dans les Collèges 1990 - fin du cycle d'orientation.
Service :
Ministère de l'Education Nationale. Direction de l'Evaluation et de la Prospective. Sous Direction de l'Evaluation du Système Educatif. Département de l'évaluation pédagogique des élèves et des étudiants.
Disciplines évaluées :
Français, Mathématiques, Allemand (LV1), Anglais (LV1), Epreuve transdisciplinaire.
Matériel :
Cahiers de l'élève; cassette pour les LV(allemand, anglais) ; questionnaire « Vie Scolaire » ; Livret de consignes (application, codage), Questionnaire professeurs.
Résultats :
Non publiés (documents internes).

Radica, K. (1990 ?). Evaluation fin de troisième d'enseignement général 1990. Document DEP.

Harnois, J. (1991). Exploitation et analyse des résultats d'une évaluation pédagogique en fin de troisième technologique réalisée en avril 1990. Rapport de stage CGSA.

Evaluation 1990 Troisième technologique :

Titre :
Evaluation dans les Collèges 1990 – fin du cycle d'orientation
Service :
Ministère de l'Education Nationale. Service de la Prévision, des Statistiques et de l'Evaluation. Direction des lycées et collèges. Division de l'Evaluation du Système Educatif. Département de l'évaluation dans les collèges et les lycées.
Disciplines évaluées :
Français, Mathématiques, Anglais (LV1), Sciences et technologies industrielles, biologiques et sociales, tertiaires ; épreuve transdisciplinaire.
Matériel :
Cahiers de l'élève et feuilles de réponse mécanographique ; cassette pour les LV ; questionnaire « Vie Scolaire » ; Livret de consignes (application, codage) ; questionnaire enseignant.
Résultats :
Non publiés (documents internes).

Evaluation fin de troisième d'enseignement technologique 1990. document DEP, non daté, non signé.

Evaluation 1995 :

Titre :
Evaluation en classe de troisième générale et technologique.
Service :
Ministère de l'Education Nationale. Direction de l'Evaluation et de la Prospective.
Disciplines évaluées :
Français, Mathématiques, Allemand (LV1 et LV2), Anglais (LV1 et LV2), Espagnol (LV1 et LV2), Sciences de la Vie et de la Terre, Physique et Chimie, Histoire - Géographie, Technologie.
Matériel (versions différentes pour les troisièmes générales et technologiques) :
Cahiers de l'élève; cassette pour les LV(allemand, anglais) ; questionnaire « Vie Scolaire » ; Document à l'attention du professeur (consignes d'application, codage, questionnaire enseignant).
Résultats :

(1996). Evaluation en classe de troisième générale et technologique 1995, premiers résultats, par discipline. Document provisoire - août 1996. (une brochure par discipline).

(1997) Evaluation pédagogique en fin de troisième générale et technologique 1995. Les dossiers d'Education & Formations, 86, mai 1997.

Evaluation 1999 :

Titre :
Etude comparative en classe de troisième.
Service :
Ministère de l'Education Nationale de la Recherche et de la Technologie. Direction de la Programmation et du Développement.
Disciplines évaluées :
Mathématiques, Sciences de la Vie et de la Terre, Physique et Chimie.
Matériel :
Cahiers de l'élève; Document à l'attention du professeur (consignes d'application, codage, questionnaire enseignant).
Résultats :
Non publiés (documents internes).

Dubreux, A. (2000 ?) Evaluation du niveau des élèves en fin de troisième et comparaison avec l'évaluation de 1995. Rapport de stage de A. Dubreux (Magistère de modélisation appliquée, Paris 10).

(2000). Troisième 1995-1999. Niveaux en mathématique, en biologie et en physique croisés. Note technique interne.

(2000). [Dépouillement du questionnaire « vie scolaire ».] Présentation et premières constatations. Note technique interne

(2000). Approche pédagogique des sources d'échec. Note technique interne

(2000). Evaluations fin de troisième 1995 et 1999. Note technique interne.

Autres évaluations :

Evaluation en Arts Plastiques :

Levasseur, J., & Shu, L. (1998). Compétences en arts plastiques des élèves de troisième de collège. Evaluation réalisée en mai-juin 1997. Paris : DPD.

Evaluation des compétences en communication :

Levasseur, J., & Trussy, C. (1997). Parler en allemand, en anglais en français. *Les dossiers d'Education & Formations*, *91*, septembre 1997.

Informatique :

IGEN (2000). Enseignement de mise à niveau en informatique en classe de seconde. Etat des lieux et propositions. 2000-005. janvier 2000. Ministère de l'éducation nationale de la recherche et de la technologie.

IGEN (2001). Enseignement de mise à niveau en informatique en classe de seconde. 2001-008. février 2001. Ministère de l'éducation nationale de la recherche et de la technologie.

Publications hollandaises :

CITOGROEP (2000). Periodieke Peiling van het Onderwijsniveau (PPON). Arnhem : Author.

Publications allemandes :

Landesinstitut für Schule und Weiterbildung. Nordrhein-Westfalen.

Publications anglaises :

Qualification and Curriculum Authority, Dfee. Mathematics test, 1999.

III - CONCLUSION

Nous l'avons vu, le Brevet est l'examen qui sanctionne la fin des études au collège et qui, depuis 1962, représente en fait le diplôme de fin de scolarité obligatoire.

Depuis 1984, des dispositifs d'évaluation de bilan périodique ont été mis en place pour suivre l'évolution du niveau des jeunes en fin de collège, et fournir aux responsables du système éducatif des outils objectifs de pilotage. Ces informations sont par ailleurs indispensables pour alimenter le débat public.

Ces deux outils sont, chacun à sa place, nécessaires et nous proposons donc de les maintenir en les améliorant pour qu'ils répondent bien, l'un et l'autre, aux objectifs affichés : information sur l'état du système éducatif, et certification.

Pour ce qui touche à la certification, il nous semble capital qu'à l'issue de la classe de 3ème, au moment où les jeunes quittent le collège et achèvent leur formation commune de base pour poursuivre leur scolarité dans des voies très diverses (souvent très spécialisées), une première certification intervienne, qui permette de garantir que les éléments de base de la culture de « l'honnête homme » du XXIème siècle sont en place : il y a là une nécessité individuelle, car la bonne intégration dans notre société est à ce prix ; et une nécessité collective, car « l'honnête homme du XXIème siècle » est d'abord un citoyen, et neuf années de scolarité obligatoire sont censées le garantir.

Pour toutes ces raisons, il nous semblerait dommageable d'attendre que le jeune ait 18, 19 ou 20 ans pour qu'intervienne la première certification de fin d'études générales ou professionnelles : brevet d'études professionnelles ou baccalauréat.

C'est d'ailleurs ce qui se fait dans la plupart des pays européens et nous voyons mal que la France renonce à cette pratique, au moment où l'intégration européenne franchit une nouvelle étape significative.

Les évaluations-bilans périodiques sont tout aussi nécessaires en fin de 3ème car c'est le point de la scolarité où se trouvent encore réunis la totalité des jeunes d'une génération, et c'est donc le moment le plus opportun pour suivre l'évolution d'ensemble de leur niveau. Elles constituent, de ce fait, un moyen irremplaçable de pilotage du système aux différents échelons : ministériel, régional, local, en permettant au Ministre et aux responsables de mettre en œuvre les ajustements politiques, techniques et pédagogiques qui aideront chacun à progresser.

ANNEXE 1 : L'EPREUVE DE FRANÇAIS

Le DNB a été modifié en 1999 (note de service n° 99-123 du 6 septembre 1999, publiée au BO n° 31 du 9 septembre 1999). On évoquera donc l'épreuve telle qu'elle existait avant cette note de service, sa forme actuelle, les connaissances et les compétences qu'elle permet d'évaluer, son fonctionnement d'ensemble et les modifications qui sont envisageables.

1 - EVOLUTION

- **Etat antérieur**

L'ancien Brevet proposait aux élèves un texte suivi de questions et débouchait sur deux sujets de rédaction, au choix du candidat. Une dictée permettait une évaluation dans le domaine de l'orthographe.

Cette formule avait trouvé ses limites, pour trois raisons essentielles :

- Le système des questions, qui isolait les rubriques « grammaire », « vocabulaire » et « compréhension » n'était plus dans la logique de programmes qui mettent en avant le principe du décloisonnement et envisagent de façon complémentaire l'ensemble des sous-disciplines qui constituent la discipline « français ». La grammaire, le vocabulaire, l'orthographe ne sont plus envisagés isolément les uns des autres mais considérés comme autant d'éléments mis en œuvre conjointement dans la lecture et/ou l'écriture d'un texte.

- Elle s'était de plus très vite figée et sclérosée, avec, en particulier, des systèmes de questionnement répétitifs et mécaniques. En vocabulaire, on faisait relever des champs lexicaux et étudier les préfixes et les suffixes ; en grammaire, on en était resté à des questions de transformation quasiment mécanique (passer de la voix active à la voix passive, du discours direct au discours indirect, de la cause à la conséquence …) qui étaient un reste des années structuralistes et ne présentaient aucun intérêt dans l'évaluation et donc la formation des compétences langagières des élèves. Il est à remarquer que cette dérive était beaucoup plus liée aux conditions d'élaboration des épreuves (sujets académiques) qu'à leur définition même, qui appelait des questions visant à évaluer les compétences des élèves dans le domaine de la compréhension des textes.

- Elle permettait une spécialisation abusive dans la préparation des sujets de rédaction. Le choix étant laissé au candidat entre un sujet dit « d'imagination » - qui amenait surtout à raconter - et un sujet dit « de réflexion » - qui demandait la mise en œuvre de diverses formes d'argumentation. Les professeurs privilégiaient l'imagination dans les classes les plus faibles et la réflexion dans les classes d'un bon niveau.

Dans ces conditions, les enseignants de collège se trouvaient placés devant un dilemme que certains ne savaient pas résoudre : fallait-il préparer les élèves au Brevet ? Fallait-il les préparer prioritairement à la classe de seconde ?

- **Etat actuel**

Il est à noter que dès les années 95 et 96, de nouveaux sujets, qui privilégiaient le décloisonnement, sont apparus dans certaines académies, à l'initiative des IA-IPR et des professeurs chargés de l'élaboration des sujets. Ils ne correspondaient pas totalement à la définition des épreuves mais n'étaient pas non plus en contradiction avec elles… Le mouvement est donc parti des académies et il est important de le noter.

Dans la logique des nouveaux programmes, appliqués à la rentrée 98 en classe de troisième, les épreuves ont ensuite évolué dans trois directions essentielles :

- Les questions posées sur le texte ne sont plus réparties entre les rubriques « grammaire », « vocabulaire » et « compréhension ». Elles se proposent d'évaluer la capacité des élèves à construire le sens d'un texte et les sous-disciplines citées concourent à cette élaboration qui se trouve naturellement placée sous le signe du décloisonnement.

- Les questions grammaticales et lexicales essentiellement mécaniques disparaissent. L'ensemble des questions est orienté vers la compréhension du texte.

- Un seul sujet de rédaction est proposé, de façon à éviter les spécialisations précoces.

- L'évaluation de l'orthographe est diversifiée et renforcée. Une dictée, un exercice d'orthographe, des questions, permettent des approches complémentaires, auxquelles s'ajoute la prise en compte renforcée de la correction orthographique dans l'exercice de rédaction.

De ce fait, les enseignants préparent leurs élèves à la classe de seconde en les préparant au Brevet. Certes, le niveau d'exigence n'est pas le même, mais l'esprit des épreuves est le même que celui de l'enseignement en classe de seconde.

- **Choix dans le domaine de l'évaluation**

Les épreuves actuelles évaluent les compétences que l'élève est censé acquérir en français pendant les quatre années qu'il passe au collège ; elles se situent absolument dans la logique de ces apprentissages préalables.

Plus précisément, on soulignera les points suivants :

- La maîtrise de la langue est évaluée, dans les trois domaines essentiels de l'orthographe, de la syntaxe et du lexique.

- L'aptitude à comprendre un texte est évaluée par l'intermédiaire des questions qui accompagnent le texte support.

- L'aptitude à s'exprimer clairement est évaluée par l'intermédiaire des réponses aux questions et de la rédaction.

- La rédaction, telle qu'elle est définie par le texte réglementaire (*« Un sujet de rédaction (…) amène le candidat à produire un texte mettant en œuvre une ou plusieurs des formes de discours étudiées au collège. La situation de communication dans laquelle doit s'inscrire le texte à produire est indiquée dans le sujet* ») permet d'évaluer la capacité du candidat à s'exprimer en fonction d'une « situation de communication », c'est-à-dire en s'adressant à une personne précise pour, par exemple, la convaincre, la persuader, l'amuser ou l'émouvoir, l'apitoyer ou susciter en elle l'indignation. Cette approche nouvelle prépare directement les élèves à l'« écriture d'invention », telle qu'elle se développe au lycée.

Au total, on peut donc estimer que le DNB, tel qu'il vient d'être modifié, en français permet l'évaluation des compétences essentielles attendues chez un élève de seconde. Des améliorations seraient cependant possibles. Elles seront exposées dans la dernière partie de cette note.

2 - FONCTIONNEMENT

On s'en tiendra, pour cette partie à deux remarques, la première « apparemment de détail », la seconde de fond.

« L'usage d'un dictionnaire est autorisé… »

Le texte réglementaire précise : « *L'usage d'un dictionnaire de langue française est autorisé* », en dehors bien sûr de la partie de l'épreuve consacrée à la dictée. C'est là la simple prise en compte d'une très ancienne demande des enseignants … Il s'est avéré, l'an passé, que, dans certains cas, il était très difficile, voire impossible, d'obtenir que les établissements prêtent un dictionnaire aux élèves les plus défavorisés. Il a même été précisé dans certaines académies que cet usage n'était pas obligatoire. Une solution rapide doit être apportée à ce problème.

L'élaboration des sujets

C'est dans ce domaine que les difficultés existent. Si un mode d'élaboration clair n'est pas défini, si les académies avancent en ordre dispersé, on observera des écarts importants, on retrouvera très vite les dérives déjà constatées par le passé. Ce phénomène sera particulièrement inacceptable à partir du moment où l'obtention du Brevet conditionnerait le passage en classe de seconde.

On peut imaginer deux dispositifs :
- élaboration, sous la responsabilité de l'Inspection Générale, et en liaison avec la DESCO, d'un sujet national,
- élaboration, sous la responsabilité de l'Inspection Générale et en liaison avec le DESCO, d'un nombre limité de sujets inter académiques.

Dans les deux cas, il sera nécessaire de réunir les concepteurs de sujets et de suivre l'élaboration de ceux-ci.

3 - MODIFICATIONS

Deux hypothèses peuvent être avancées :

- Evaluer les compétences des élèves dans le domaine de l'oral, par exemple sous la forme d'un contrôle en cours de formation, et en liaison, en $4^{ème}$, avec les travaux croisés.

- Relier plus étroitement les épreuves du Brevet au programme de la classe de troisième et proposer un texte emprunté aux genres qui doivent être étudiés à ce niveau, à savoir :
  - Une œuvre à dominante argumentative,
  - Une œuvre autobiographique française,
  - Un ensemble de textes poétiques du XIXème ou du XXème siècle,
  - Une pièce de théâtre du XIXème ou du XXème siècle,
  - Deux romans ou un roman et un recueil de nouvelles du XIXème ou du XXème siècle.

Il serait possible, de la sorte, de favoriser la constitution d'un socle culturel commun, exigible à l'entrée en seconde.

Katherine **Weinland**, IGEN, groupe des Lettres

ANNEXE 2 : L'EPREUVE DE MATHEMATIQUES

Comme pour l'ensemble des disciplines, la partie de l'examen concernant les mathématiques est décrite dans l'arrêté du 18 août 1999 et dans la note de service du 6 septembre 1999, parue au B0 n° 31 du 9 septembre 1999. La note de mathématiques est obtenue à partir d'un contrôle ponctuel final (coefficient 2) et d'une note de contrôle en cours de formation pour les classes de quatrième et de troisième (coefficient 1).

Adéquation entre épreuves ponctuelles finales et programmes.

Le programme n'étant pas uniquement rédigé dans la finalité de l'évaluation ponctuelle finale. La note de service du 9 septembre précise les compétences à évaluer : "*Les acquisitions à évaluer ont pour référence les programmes des classes de troisième correspondant aux différentes séries du diplôme national du brevet ; vis-à-vis de ces programmes, elles se situent exclusivement dans le cadre des "compétences exigibles" pour la série "collège", des "capacités exigibles" pour la série "technologique" et des "compétences exigibles du référentiel du CAP" pour la série "professionnelle".*

Ces acquisitions à évaluer s'organisent autour des pôles suivants :

- *exécuter et exploiter un calcul, un graphique ou une figure géométrique ;*
- *interpréter graphiquement une situation numérique et interpréter numériquement une situation graphique ou géométrique ;*
- *mobiliser et mettre en œuvre des connaissances et des méthodes pour la résolution de problèmes simples."*

On va donc avoir des parties du programme qui seront régulièrement évaluées et d'autres systématiquement absentes. Il est difficile de tenir un discours général sur ces aspects, étant donné que les épreuves sont inter académiques, voire académiques, et qu'il n'est pas sûr que les pratiques sont uniformes d'une académie à l'autre.

Voilà par exemple, ce qu'écrivent les IA-IPR de Versailles sans que l'on soit sûr que ce propos est valide pour l'ensemble des académies.

La situation en Ile-de-France.
Une étude des notions abordées dans les sujets du brevet des collèges en Ile de France (académies de Paris, Créteil, et Versailles) au cours des sessions de juin 1996 à 2000 montre que certains points du programme de 3e sont régulièrement évalués.

Il en est ainsi de la trigonométrie dans le triangle rectangle, de la propriété de Thalès, du calcul littéral (développement et factorisation, identités remarquables), des équations « produit », des systèmes d'équations linéaires à deux inconnues, toujours introduites dans un exercice à support concret, du calcul portant sur les radicaux, des fonctions linéaire et affine.

Chaque année une question fait intervenir la propriété de Pythagore ou sa réciproque, bien que celle-ci ne figure pas explicitement au programme de 3e ; mais il est vrai qu'elle est beaucoup travaillée en classe de 4e.

Des calculs faisant intervenir des fractions sont demandés chaque année, le résultat étant d'ailleurs attendu sous la forme « la plus simple possible ». Que la notion de *fraction irréductible* soit maintenant au programme de la classe de 3e devrait permettre de formuler plus directement et explicitement la question[6].

[6] On peut cependant s'interroger sur la pertinence de telles questions à une époque où les calculatrices « bas de gamme » donnent le résultat sous forme de fraction irréductible. Il ne semble d'ailleurs pas y avoir d'apprentissage en classe sur cette utilisation.

Certaines notions ne sont pratiquement pas évaluées. On peut citer notamment, en géométrie, ce qui tourne autour des vecteurs, des polygones réguliers, des sections planes de solides. Le cas de « angle inscrit » est un peu particulier, puisqu'il ne faisait pas partie des notions exigibles en fin de 3e jusqu'en 1998 inclus.

Dans la partie « travaux numériques », on peut constater que la résolution d'inéquations du premier degré à une inconnue n'est pratiquement pas demandée.
La partie « statistiques descriptives » du paragraphe C – Gestion de données et fonctions du programme de 3e n'est jamais évaluée. Cause ou conséquence, on constate que son apprentissage est réduit dans les pratiques de classe. On peut d'autant plus le regretter que cela faciliterait l'articulation avec le programme de seconde.

La situation dans l'académie de Grenoble.

Dans cette académie, les statistiques sont systématiquement évaluées, car le parti à été pris de mettre un peu de statistiques chaque année.

À rajouter dans la liste des notions non évaluées : transformation de figures par rotation, compositions de symétries centrales et de translations, grandeurs composées, changement d'unités dans les grandeurs composées, exemples simples d'algorithmes.

Avec la même remarque que pour l'Ile-de-France : est-ce qu'on est sûr que ces trois dernières notions sont enseignées ?

En résumé, il y a un écart certain entre les programmes et les notions évaluées, même si on retrouve un certain nombre de points communs d'une académie à l'autre, les choix académiques sont nettement perceptibles.

Les épreuves sont-elles prédictives de la réussite ?

Là encore, il est difficile de tenir un propos général et valide sans qu'il soit étayé par des analyses fines.

Pour un ensemble d'élèves donnés, nous disposons de trois séries de notes : la note de contrôle continu, la note de l'écrit du brevet et la note du premier trimestre de la classe de seconde. Le travail fait sur un ensemble de classes dans l'académie d'Aix-Marseille montre qu'il y a, pour un groupe d'élèves donné provenant de la même classe ou du même collège, peu d'écart entre les moyennes, et que les différents coefficients de corrélation calculés sont autour de 0,7, ce qui n'est pas étonnant pour une étude avec de tel biais :

- on travaille sur de petits effectifs ;
- le contrôle continu est toujours influencé par le contexte local ;
- l'élève évolue au cours du temps (heureusement !)

Dans l'académie de Clermont-Ferrand, les notes de début de seconde sont toujours bien plus basses que celles du brevet des collèges. Pourtant les sujets permettent de faire le point sur une grande partie du programme de troisième. Mais les exigences des professeurs de seconde se situent à d'autres niveaux : rapidité, contenu plus technique.

Dans l'académie de Grenoble, sur un échantillon de 190 données : contrôle continu, brevet, notes de seconde on obtient les coefficients de corrélation suivants :

- entre contrôle continu et brevet : 0,54
- entre contrôle continu et seconde : 0,73
- entre brevet et seconde : 0,58

A part le 0,73, ces coefficients sont faibles et demanderaient certainement des études plus poussées.

La note de service du 6 septembre 1999 parue au BO n°31 du 9 septembre 1999 est-elle connue et appliquée ?

En ce qui concerne l'harmonisation des corrections des épreuves d'examen, la note stipule qu'"*il est recommandé :*

- *d'organiser des réunions des correcteurs pour un échange de vue après analyse d'un premier lot de copies ;*
- *de mettre en place auprès du recteur, pendant la durée de correction des copies, une cellule comprenant des membres de la commission de choix des sujets, afin de donner toutes indications nécessaires aux correcteurs en réponse aux problèmes éventuels posés.*

Le jury vérifie l'application des barèmes et des recommandations définis par la commission académique de choix des sujets."

Les pratiques dans ce domaine semblent variables.

Concernant l'annexe 1 qui précise les épreuves de l'examen et particulièrement celle de mathématiques, il n'est pas sûr que les professeurs la connaissent, et même s'ils en ont pris connaissance, il n'est pas sûr qu'ils considèrent qu'elle apporte des indications susceptibles de leur faire modifier leurs pratiques habituelles.

Marc **FORT**, IGEN, groupe de Mathématiques

ANNEXE 3 : L'EPREUVE D'HISTOIRE-GEOGRAPHIE ET D'EDUCATION CIVIQUE

1 - Résultats d'une enquête pilotée par l'Inspection générale sur les résultats des élèves au DNB

1 1- Les justifications de l'enquête

Conjonction de la mise en place **de nouveaux programmes** et **de nouvelles épreuves au DNB** dans les trois disciplines à la rentrée 1999.

Les épreuves anciennes ne donnaient plus satisfaction :
- inadaptation des questions de cours, largement délaissées par les élèves (jusqu'à 30 % dans certains collèges) ;
- part réduite accordée à l'éducation civique : présence non obligatoire ou sujets souvent marginaux et mal adaptés aux démarches pédagogiques de cette discipline.

Mise en place, à la session 2000, d'épreuves plus proches des pratiques de classe et mieux ciblées sur l'évaluation de compétences fondamentales au collège, disciplinaires ou transdisciplinaires :
- *maîtrise des connaissances fondamentales en histoire, géographie et éducation civique ;*
- *aptitude à lire et à mettre en relation des documents ;*
- *aptitude à rédiger et à argumenter ;*
- *maîtrise de la langue.*

Les objectifs de cette enquête étaient donc d'observer les résultats des élèves pour évaluer le degré d'acquisition des compétences fondamentales à la fin du 1er cycle, mais aussi de porter un regard sur la validité des exercices proposés dans le cadre des nouvelles épreuves du DNB.

1.2- Une enquête approfondie portant sur 15 000 copies d'élèves

Dans 8 académies (Caen, Lille, Montpellier, Nantes, Nice, Rennes, Rouen et Strasbourg), les IA-IPR d'histoire et géographie, en collaboration avec les IA-DSDEN, ont conduit, dans un ou plusieurs départements, une enquête auprès des professeurs correcteurs.

Cette enquête a été conduite « à chaud », au moment des corrections, à partir d'un questionnaire commun comportant trois volets :
- un relevé quantitatif des résultats obtenus à chacun des exercices proposés dans le cadre de l'épreuve ;
- une analyse qualitative du niveau de maîtrise par les élèves des compétences fondamentales évaluées ;
- un avis des correcteurs sur chacun des sujets proposés (choix et libellé des sujets ; choix des documents ; clarté, formulation, niveau de difficulté des questions…).

2. Quelques constats majeurs

2.1. Des résultats globalement encourageants

La mobilisation des candidats a été nettement supérieure à celle des autres années (très faible taux d'exercices non traités (5 %).

La moyenne générale est, dans la grande majorité des académies, supérieure à celle des années précédentes (souvent de l'ordre de 1 point sur 20).

Mais on observe aussi le maintien d'une grande disparité de résultats entre les collèges :

Exemple : département de Loire-Atlantique, option LV2 (130 collèges) :

- - moyenne du département : 11/20
- - collège ayant la moyenne la plus élevée : 13/20
- - collège ayant la moyenne la plus faible : 7,2/20

NB : une remarque qui ne manque pas d'intérêt : les observations mettent en évidence, dans la majorité des cas, une nette corrélation dans les résultats obtenus par les candidats en français et en histoire-géographie-éducation civique.

2.2. Des niveaux de réussite inégaux aux différentes parties de l'épreuve

Niveau de réussite pour les différents critères retenus :
(++) : réussi par la majorité des élèves
(+) : réussi par environ la moitié des élèves
(-) : réussi par une minorité d'élèves.

Lecture et mise en relation de documents ***(réponse à des questions)***	**Rédaction d'un paragraphe argumenté** ***(20 lignes en histoire-géographie ; 15 lignes en éducation civique*****)**	**Repères chronologiques et spatiaux** ***(portant sur les quatre niveaux du collège)***
- Compréhension des questions **(++)** - Prélèvement des informations **(++)** - Compréhension globale des documents **(+)** - Mise en relation des documents **(-)**	- Compréhension et respect du sujet **(+)** - Exactitude des connaissances **(+)** - Pertinence des informations tirées des documents **(++)** - Pertinence des informations personnelles complémentaires **(-)** - Cohérence de l'argumentation **(-)** - Qualité de la rédaction **(+)** - Longueur conforme au texte de l'épreuve (+)	- Taux global de réussite assez médiocre - Maîtrise plus faible des repères portant sur les programmes de 6ème et 5ème - Meilleure maîtrise des repères spatiaux que des repères chronologiques

Remarques concernant la lecture et la mise en relation de documents :

- le prélèvement pertinent d'informations dans les documents est globalement maîtrisé ;
- la difficulté principale porte sur le croisement d'informations dans plusieurs documents.

Remarques concernant le paragraphe argumenté :

C'est la partie la plus nouvelle de l'épreuve ; les difficultés rencontrées par les élèves sont de trois ordres :

- cohérence de l'argumentation : difficulté à bâtir un raisonnement logique (en utilisant, notamment, les mots de liaison servant à l'argumentation) ;
- faiblesse de l'apport de connaissances personnelles extérieures aux documents ;
- difficulté à passer du particulier au général, à s'appuyer sur le cours pour généraliser à partir des exemples présentés dans les documents. En éducation civique, difficulté particulière à passer des études de cas au texte de référence, à la mise en relation du vécu avec les valeurs fondamentales de la société.

2.3. L'importance du libellé des sujets

L'enquête a montré l'importance du choix des sujets et de leur libellé dans les résultats des élèves

En éducation civique, en particulier, l'utilisation de notions telles que « opinion publique », « valeurs et symboles de la République » dans le libellé des sujets, a constitué un obstacle pour un nombre important d'élèves.

3. Quelques perspectives

Même si leur mise en œuvre mérite quelques ajustements, **les nouvelles épreuves d'histoire-géographie et d'éducation civique donnent globalement satisfaction ; elles permettent une évaluation relativement fine et pertinente du niveau d'acquisition des compétences exigibles à la fin du collège dans ces disciplines.**

Pour les prochaines sessions, **un suivi national** serait utile pour une plus grande convergence des sujets et de leur formulation.

L'enquête, conduite pour la session 2000, mérite d'être complétée et approfondie. Pour la session 2001, l'Inspection générale pilote une nouvelle enquête qui porte sur 10 académies et qui s'élargit dans deux directions importantes :

- extension à l'analyse des résultats des élèves des séries technologiques et professionnelles (pour lesquelles la situation s'avère très différente) ;
- mise en relation, dans les académies concernées, des résultats obtenus aux épreuves du DNB et des résultats obtenus aux tests d'entrée en seconde.

D'ores et déjà, les résultats de cette enquête donnent des indications intéressantes sur **un certain nombre de compétences inégalement maîtrisées en fin de troisième, et sur lesquelles il convient de mettre l'accent tout au long du collège**:

- la confrontation de plusieurs documents ;
- l'écrit : exercices réguliers de rédaction et travail sur l'argumentation, conduisant progressivement à la réalisation d'un paragraphe argumenté ;
- la maîtrise des notions fondamentales et des repères chronologiques et spatiaux, en nombre restreint, qui constituent l'armature indispensable à la culture commune de tous les élèves à la fin du 1er cycle.

Michel **HAGNERELLE,** IGEN, Groupe histoire-géographie, éducation civique

ANNEXE 4 : EPREUVES STANDARDISEES ET BREVET :

1 - Pourquoi introduire des épreuves standardisées ?

Les études de la notation au brevet ont montré qu'il existe des différences importantes dans les notations indépendantes effectuées sur la performance de l'élève. Ces effets sont particulièrement massifs dans le contrôle continu. Comment pallier ces inconvénients ? Les recherches docimologiques qui ont, depuis longtemps (Piéron, 1963), pour toutes les situations de notation, mis en évidence des biais importants dans les notations et les examens, concluent invariablement à l'intérêt du recours à des épreuves standardisées.

Les résultats obtenus à des épreuves standardisées nationales, pourront servir de base pour mettre en place des procédures d'harmonisation des notations. Signalons que plusieurs pays (Grande-Bretagne, Suède) recourent à de telles procédures, dites de "modération". L'intégration d'une telle épreuve introduira un facteur correctif aux écarts de notation à niveau égal.

La mise en place d'épreuves standardisées permettra, également, de garantir l'équivalence des sessions successives (les banques d'items seraient ici d'une aide précieuse).

2 - Qualités de ces épreuves

Pour construire ces épreuves, il sera nécessaire de définir le programme qu'elles doivent couvrir et leur niveau de difficulté.

Nous insistons notamment sur deux points : (1) ces épreuves doivent avoir des qualités psychométriques solides ; (2) il s'agit d'une évaluation certificative, cette épreuve ne peut donc servir d'épreuve d'évaluation diagnostique pour une pédagogie différenciée en seconde, encore moins d'épreuve d'évaluation du niveau national (sauf à rester dans les erreurs dénoncées sur le rapport entre objectifs et nombre d'items par exemple).

3 - Veiller à leur acceptation

Il est important de s'interroger sur l'accueil que recevra une telle mesure dans la société. En effet, malgré les lourdes critiques souvent formulées à l'encontre des notations et des examens, force est de constater la persistance de ce système. Mis à part les fameux QCM utilisés dans les études de médecine, et l'introduction d'épreuves standardisées de type plus ou moins psychométriques, pour l'entrée dans certaines écoles de commerce et d'ingénieurs, ou dans des formations paramédicales, les épreuves des examens restent largement des épreuves de forme scolaire. L'introduction d'épreuves standardisées devra être discutée et argumentée avec les différents acteurs concernés, surtout si elle conduit à l'abandon progressif d'épreuves traditionnellement jugées, par certains, plus aptes à mettre en évidence les qualités d'un élève (par exemple la dissertation).

4 - Le brevet : examen ou bilan ?

Enfin, on peut se demander si dans ce mouvement de réorganisation profonde du brevet, il ne serait pas utile d'aller plus loin en adoptant un système de « niveaux » en plus d'un système d'examen en « tout ou rien ». Comme le montre la docimologie, placer une coupure dans un continuum est sans doute le point le plus problématique du traitement de résultats d'examen. Etablir une coupure renvoie nécessairement à la problématique de la compétence minimale toujours difficile à gérer.

Nous voyons plusieurs arguments en faveur d'une autre solution : celle de la définition de niveaux de compétences. Ce mode de fonctionnement est plus en accord avec le monde professionnel où l'on utilise les compétences par niveaux. Nous avons rencontré ce type de fonctionnement dans le passeport informatique, et les diplômes de langues à usage professionnel. Pour un recrutement, on établit un profil de poste et on ne demande pas que toutes les compétences soit au niveau le plus élevé. Par exemple, en

informatique, de nombreux postes requièrent de savoir utiliser un traitement de texte, et l'on ne demande pas forcément de savoir installer et dépanner les logiciels. Dans cette même optique, un brevet qui fonctionnerait par niveaux aurait une utilité beaucoup plus importante, pour la vie professionnelle, qu'un brevet en « tout ou rien ».

Nous voyons également un autre avantage à ce système d'évaluation : placer une coupure veut dire que certains échoueront – de plus, ceux qui échoueront sont ceux qui risquent fort de ne pas obtenir de diplôme de niveau plus élevé – alors que ce sont justement ceux là qui auront besoin d'utiliser ce diplôme à leur entrée dans la vie professionnelle. Une solution consistant à faire du brevet un premier bilan qui reconnaît les compétences acquises dans les différents domaines, serait certainement plus utile socialement. Le brevet deviendrait ainsi une première étape pour l'évaluation et la formation tout au long de la vie.

La construction d'un portefeuille de compétences serait sans doute une bonne démarche de préparation à l'identification des compétences dans une optique professionnelle. Dans le cadre du bilan de compétences, de la validation des acquis, il est devenu essentiel pour la personne, de faire la preuve qu'elle possède telle ou telle compétence. La démarche de constitution d'un portefeuille de compétences repose sur la collecte de documents significatifs, que la personne peut ensuite utiliser et valoriser auprès des différents acteurs pour établir ses compétences. Signalons au passage qu'un certain nombre de séquences pédagogiques mises en place par les Conseillers d'Orientation-Psychologues ont des objectifs voisins de ceux de la démarche de constitution d'un portefeuille de compétences. Nous plaiderons ici en faveur de l'intérêt d'introduire ce type de démarche lors du brevet. Par exemple, pour l'épreuve orale, pourquoi les élèves ne viendraient-ils pas présenter un portefeuille de leurs travaux de collège, voire de primaire, collectés dans une ou plusieurs disciplines ?

5 - Des indicateurs sur le rendement des établissements

Dans la mesure où on s'intéresse au rendement des établissements, des épreuves standardisées, plus fidèles et faciles à interpréter que les notes, seraient de meilleurs indicateurs de la performance des élèves en fin de troisième. Comme tous les élèves subissent en sixième les épreuves d'évaluation prévues pour ce niveau, on disposerait ainsi, pour chaque élève, d'un indicateur sur ses acquis à l'entrée et à la sortie du collège. La mise en relation de ces deux indicateurs devrait être, dans un premier temps, maniée avec prudence. En effet, les deux épreuves ne sont pas de même nature et ne sont pas administrées dans les mêmes conditions. Les erreurs de mesure sur chacune des épreuves et plus encore sur les différences entre les protocoles risquent d'être importantes.

ANNEXE 5 : L'EVALUATION DES ACQUIS DES ELEVES A LA FIN DES CYCLES D'APPRENTISSAGE (Extraits du rapport de l'IGEN, 1991, chapitre 9)

La notion d'évaluation est désormais consacrée par les textes de la loi d'orientation de 1989 et par les nouveaux statuts des professeurs.

Jusqu'alors, les textes officiels ne mentionnaient même pas l'obligation faite aux enseignants de noter les copies des élèves, si ce n'est lors des compositions trimestrielles, comme le précisait l'arrêté du 7 juillet 1890. Il, fut nécessaire de recourir à l'interprétation jurisprudentielle.

C'est la loi d'orientation qui, pour la première fois, prescrit aux enseignants l'obligation d'évaluer les acquis des élèves : « les enseignants apportent une aide au travail personnel des élèves et en assurent le suivi. Ils procèdent à leur évaluation. Ils les conseillent dans le choix de leur projet d'orientation en collaboration avec les personnels d'éducation et d'orientation. »

Les décrets statutaires du 18 septembre 1989 des professeurs agrégés, certifiés, d'éducation physique et sportive et de lycée professionnel, précisent également, pour la première fois, et selon une rédaction commune que « dans le cadre [de leur service d'enseignement] ils assurent le suivi individuel et l'évaluation des élèves et contribuent à les conseiller dans le choix de leur projet d'orientation. »

Il a donc paru opportun d'étudier comment, sous ses diverses formes, traditionnelles ou nouvelles, l'évaluation se trouvait pratiquée au sein du système éducatif et de s'efforcer de faire, en quelque sorte, une évaluation de l'évaluation.

Opération intellectuelle complexe, l'évaluation s'inscrit, à son heure, au nombre des obligations des enseignants et du système éducatif tout entier, voire du service public dans son ensemble.

L'École doit aux familles les indications précises et utiles qui permettront aux élèves de savoir quelles connaissances et quelles méthodes ils maîtrisent et celles qui leur font défaut, les voies par lesquelles ils peuvent progresser et ce à quoi ils peuvent aspirer.

C'est d'ailleurs une insuffisance de notre système éducatif que d'identifier difficilement les innovations et, par conséquent, de mal savoir diffuser celles qui peuvent se révéler porteuses d'avenir.

Une esquisse d'état des lieux

Les thèmes d'investigation choisis

Pour tenter d'appréhender, dans sa globalité comme dans sa diversité, la réalité des pratiques d'évaluation, on a choisi de l'observer en fonction des catégories qu'utilisent traditionnellement les spécialistes dans ce domaine.

Ce n'est donc pas par hasard que presque tous les sujets d'investigation choisis recoupent ces catégories. Les évaluations « sommative » - bilan des connaissances et méthodes acquises à un moment donné - et « formative » - saisie d'information sur le degré de maîtrise des connaissances en cours d'acquisition, en vue de permettre au maître d'adapter son enseignement et à l'élève de travailler plus efficacement - recoupent le thème des pratiques d'évaluation des professeurs en classe de troisième et de seconde. L'évaluation « certificative » est illustrée par la prise en compte des acquis dans la certification des élèves à la fin du collège. L'évaluation « prédictive » est appréhendée dans la prise en compte des acquis pour l'orientation des élèves en fin de classe de troisième.

La notation chiffrée, accompagnée ou non de l'annotation, constitue actuellement l'instrument essentiel de toutes ces formes d'évaluation.

Il en va différemment pour l'évaluation d'une autre sorte que constituent les études sur grands échantillons, menées depuis une dizaine d'années par l'administration centrale et pour l'opération nationale conduite, à la rentrée 1989-1990, sur les acquis des élèves en classes de CE2 et de sixième. La nécessité de réaliser, à la demande du ministre, une évaluation « d'accompagnement » de l'opération CE2-6è et le souci d'essayer de tirer les enseignements des enquêtes effectuées depuis dix ans, quels qu'en soient les promoteurs, ont conduit à prendre pour autre thème d'investigation « l'évaluation des acquis des élèves à la fin des cycles d'apprentissage et d'approfondissement » et de tenter de tirer « une leçon des évaluations » antérieures.

Les champs d'observation choisis et les modalités de l'investigation ont découlé tout naturellement de la préoccupation de ne négliger aucune forme d'opération qui puisse, à un titre ou à un autre, être considérée comme d'évaluation.

L'enquête a porté, parallèlement :

- sur l'entreprise d'évaluation des acquis des élèves à la fin des cycles d'apprentissage et d'approfondissement de l'école élémentaire ; opération CE2-6e engagée en septembre 1989 ;
- sur les pratiques d'évaluation des professeurs, en classe de troisième et de seconde ;
- sur la prise en compte des acquis pour la certification des élèves à la fin du collège, ainsi que pour leur orientation.

On a, enfin, tenté d'identifier les éléments qui devraient permettre de tirer la leçon des évaluations conduites dans la dernière décennie.

Les méthodes utilisées ont été diverses, dans le souci de retenir celles qui convenaient le mieux au domaine évalué et d'explorer celles qui semblaient le mieux correspondre à la mission de l'Inspection générale et aux délais imposés. On a ainsi recouru alternativement à des entretiens individuels ou de groupe, directifs ou non, à des questionnaires ouverts ou fermés, à des collectes de documents administratifs et pédagogiques, à l'observation guidée dans les classes, dans les conseils de classes et dans les stages de formation. Les échantillons ont été importants ou très limités selon les objectifs visés, validation d'une hypothèse ou identification d'un phénomène. Ils ont été choisis tantôt dans quelques académies tantôt dans toutes. Les données recueillies ont fait l'objet d'analyses de contenu ou de traitement informatique par expérimentation d'un logiciel. On a également procédé à une analyse des travaux déjà réalisés.

Ont été associés à l'opération des élèves, des parents, des enseignants et des formateurs, des membres des corps d'inspection, des établissements publics sous tutelle, des organismes de recherche, des universitaires et des chercheurs. Les administrations centrales et rectorales ont été consultées et des organisations syndicales interrogées.

Ce souci de diversifier les champs et les modes d'investigation n'a pourtant pas permis d'appréhender la richesse de toutes les initiatives prises par les enseignants ou les responsables administratifs et pédagogiques.

L'évaluation par la note

Quelles que soient la richesse et la diversité des discours sur l'évaluation, la note constitue le moyen quasi exclusif de rendre compte des performances de l'élève, d'exprimer un diagnostic, de fonder une certification ou un pronostic.

L'évaluation des acquis en classe : notes et annotations

L'évaluation des acquis des élèves se fait dans la presque totalité des cas, à partir de devoirs écrits. Dans les quatre disciplines (français, histoire et géographie, mathématiques et sciences physiques) qui ont fait l'objet d'une enquête auprès d'un échantillon de 1 434 élèves et de 443 professeurs de classes de troisième ou de seconde, les élèves réalisent en moyenne chaque trimestre :

- 11 devoirs par discipline en classe de troisième (5 en classe, 6 à la maison) ;
- 8 devoirs par discipline en seconde (4 en classe et 4 à la maison).

En troisième comme en seconde, une part seulement des évaluations (15 %) s'appuie sur des exercices oraux. Les professeurs estiment cependant que l'appréciation globale qu'ils portent sur un élève en fin d'année prend en compte l'oral, à raison de 25 %.

Tous les devoirs sont, de l'avis des professeurs, corrigés et annotés, mais seuls les devoirs réalisés en classe et en temps limité sont systématiquement notés. Ceux qui sont faits en temps libre, à la maison, ne le sont qu'à proportion de la moitié en troisième et du tiers en seconde.

De l'ensemble de ces constats on peut conclure que :

- la prédominance de l'écrit est très forte, en troisième comme en seconde ;
- les devoirs en classe, tous notés, visent plus particulièrement à dresser le bilan de ce que l'élève sait et sait faire ;
- les devoirs à la maison ont pour objectif essentiel d'aider les élèves dans leurs apprentissages.

Les élèves ne tirent pas tout le profit escompté du lourd travail de correction que s'imposent les professeurs. Alors que la quasi-totalité des enseignants dit annoter systématiquement les copies, 30 % des élèves ont le sentiment qu'elles ne comportent pas d'annotations et 10 % déclarent que, de toute façon, ils ne les lisent pas. Moins de la moitié des élèves des collèges ou des lycées considérés comme favorisés, et moins du tiers de ceux des établissements réputés défavorisés, affirment que les commentaires portés sur les copies leur sont utiles ou très utiles.

L'explication de cette donnée est à rechercher dans la pratique des enseignants :
- deux tiers des professeurs de troisième et la moitié de ceux de seconde indiquent qu'ils apprécient les copies de leurs élèves en fonction d'objectifs bien identifiés et d'un niveau d'exigence explicite ;
- une minorité de professeurs dit pratiquer une évaluation qui inclut la mesure du progrès de l'élève : 40 % de ceux de troisième dans les collèges défavorisés, et à peine 20 % dans les lycées favorisés.

C'est parce que le progrès de l'élève n'est pas de façon assez explicite au centre des préoccupations, que ce dernier a des difficultés à interpréter notes et annotations, à apprécier les résultats de ses efforts et à mesurer les obstacles qui lui restent à franchir. Si l'on considère, d'autre part, que le travail constitue, en lui-même, un moyen essentiel de s'améliorer, il faudrait que les élèves des établissements défavorisés n'aient pas moins de devoirs à la maison que les autres. On a observé, par exemple, qu'en français, les élèves des établissements défavorisés ont un quart à un tiers de travail en moins.

En tout état de cause, la note, qui résume un nombre important d'informations, encore que variable d'une discipline à l'autre et d'un exercice à l'autre, demeure difficile à interpréter. Elle constitue cependant un instrument irremplaçable.

La notation au diplôme national du brevet

Les notes constituent également, dans le cadre de l'évaluation certificative du diplôme national du brevet, un instrument quasi exclusif. L'enquête, conduite dans trois académies, complétée par l'analyse des résultats du brevet sur plusieurs années et par des études menées, sur ce diplôme, à l'initiative de rectorats, d'inspections académiques ou de responsables, a mis en évidence une hétérogénéité importante des modes et des critères d'évaluation. Elle a également montré l'intérêt d'une meilleure exploitation des résultats des épreuves écrites.

Le diplôme national du brevet n'est plus désormais conçu comme une sanction de la scolarité obligatoire mais comme une étape dans un cursus, le total des points obtenus, plus ou moins élevé, rendant compte du degré d'acquisition des connaissances, et l'obtention de la moyenne attestant le franchissement d'un seuil.

Ce total résulte de la combinaison des notes obtenues au cours des deux dernières années dans le cadre d'un « contrôle continu » et de celles des trois épreuves de l'examen terminal. L'association originale de deux modes d'évaluation peut, dans une certaine mesure, corriger les variations de la notation, selon les établissements et les professeurs, en même temps qu'elle permet de tempérer le caractère aléatoire du seul examen. Les instructions données pour l'établissement de sujets qui prennent mieux en compte les objectifs des programmes nationaux et pour la correction des épreuves, à partir d'un barème fondé sur des critères bien définis, représentent un progrès appréciable qui n'est pas encore réalisé dans le contrôle continu. On constate ainsi que, dans les établissements dont les élèves ont un niveau relativement homogène et généralement élevé, l'examen terminal compense la grande exigence qui a présidé à la notation au contrôle continu - et qui se traduit par des notes sévères - tandis qu'il remplit une fonction inverse dans les établissements dont le niveau est hétérogène.

Des études approfondies ont permis d'observer une inégalité des taux de réussite entre les départements et les établissements, ce qui ne peut s'expliquer par les seules différences de niveau des élèves. On n'est donc pas encore parvenu à une rigueur suffisante dans l'application de critères de notation définis de façon précise, en fonction d'objectifs à atteindre et donc à une régulation satisfaisante du système de certification des acquis.

En 1989 des écarts significatifs ont été constatés entre les pourcentages de reçus : 13 points entre les académies d'Amiens et de Toulouse, 12 points entre les départements du Tarn et de l'Aveyron, 70 points entre deux établissements, 100 points entre deux classes.

Le souci de pondérer les coefficients de toutes les disciplines et d'équilibrer l'importance accordée au contrôle continu et au contrôle final n'a pas entraîné les résultats attendus. Les notes obtenues en mathématiques et en histoire et géographie à l'écrit sont apparues déterminantes pour l'obtention du brevet. A la session de 1989, la seule note de mathématiques aurait permis, par exemple dans l'académie d'Amiens, de prévoir, dans 80 % des cas, les résultats au brevet. A la session 1990, c'est la note d'histoire et géographie qui, dans la même académie, permettait de formuler un pronostic du même ordre.

Le recours à la pondération des coefficients apparaît donc comme une forme de leurre. Deux explications peuvent en être avancées :

- les variations importantes, selon les disciplines, de l'amplitude entre les notes les plus élevées et les plus basses ;
- le type d'exercice qui, dans certaines disciplines plus que dans d'autres, nécessite la mise en oeuvre de compétences plus variées et transversales », c'est-à-dire requises dans plusieurs disciplines.

Pour l'ensemble des académies, en raison de la pratique du « repêchage », 20 % des élèves sont reçus sans obtenir le total des points requis pour atteindre la moyenne de 10 sur 20. Une plus grande proportion encore des reçus n'atteint pas 10 de moyenne à l'écrit.

Ces constats conduisent à s'interroger sur la rigueur, la clarté et l'équité des résultats du brevet. La note obtenue à cet examen est en effet le résultat de l'amalgame de notes attribuées selon des modalités très différentes, parfois insuffisamment contrôlées, et qui attestent des capacités très diverses et qui ne sont pas toujours identifiées avec précision. Le fait qu'un nombre non négligeable d'élèves soit reçu sans avoir 10/20, en soi parfaitement acceptable dans la mesure où cette moyenne procède d'une construction intellectuelle relativement arbitraire, pose un réel problème de clarté pour les élèves, les parents et les professeurs, sur la localisation du seuil à franchir. La véritable question est de savoir si les procédures d'évaluation mises en oeuvre permettent d'attester que l'élève possède bien les connaissances supposées acquises à la fin du collège. L'idéal serait de pouvoir répondre à cette interrogation par oui ou par non, sans que l'on ait besoin de se référer à une note et à une moyenne. De toute façon, la variation des seuils de repêchage, au sein d'une même académie (qui permet au département dont les moyennes sont les moins élevées d'obtenir le plus grand pourcentage de reçus) pose un problème d'équité dans la certification.

Il est enfin apparu que les résultats obtenus par les élèves aux épreuves écrites du diplôme national du brevet permettent, au-delà de la simple certification, de formuler un bon pronostic de la réussite au lycée. Ce pronostic est toutefois meilleur pour les élèves des sections C d'abord, B, F et G ensuite, que pour les élèves des sections A. Il est mauvais pour les élèves de BEP.

Si les résultats du brevet ne sont pas actuellement pris en compte pour l'orientation des élèves, les notes sur lesquelles s'appuie le conseil de classe pour définir, au cours de la troisième, l'orientation des élèves, sont très précisément celles qui figurent sur les relevés du contrôle continu du brevet. Or, nous avons noté que ce ne sont pas les résultats du contrôle continu qui permettent de formuler un bon pronostic, mais les épreuves écrites, que, du fait des dispositions réglementaires, on s'est interdit d'utiliser.

L'utilisation des notes par les conseils de classe dans une perspective d'orientation

L'étude conduite dans onze collèges, choisis comme représentatifs de la variété des situations que l'on rencontre sur le territoire, a confirmé la prévalence de la note pour la décision d'orientation. Il en va ainsi dans tous les cas, que les collèges affichent ou non une politique d'orientation explicite, que les statistiques sur les flux d'orientation ou sur l'adaptation des élèves en seconde soient ou non utilisées, que les décisions soient effectivement prises dans les conseils de classe ou seulement confirmées dans cette enceinte, que les délégués des élèves et des parents participent activement ou non au débat et quel que soit le temps consacré au cas de chaque élève (de 30 secondes à 4 minutes dans les conseils considérés).

L'observation des conseils de classe fait apparaître que 50 % des interventions sont consacrées à l'information de l'ensemble des participants sur les résultats scolaires, 30 % d'entre elles concernent les comportements des collégiens et leurs attitudes face au travail. On peut aisément en déduire qu'il serait facile d'accorder davantage de temps qu'on ne le fait actuellement - 20 % des interventions – au dialogue sur les mesures qui permettraient d'aider l'élève à progresser et à l'élaboration de la décision d'orientation qui sera pour lui la plus judicieuse.

Une enquête réalisée à la fin du premier trimestre de l'année scolaire, sur 1 000 élèves admis en seconde, montre que le degré de réussite des élèves coïncide bien avec le pronostic formulé par les professeurs. 6 % seulement des élèves n'obtiennent pas de résultats conformes aux prévisions, 1 % réussissent malgré un pronostic défavorable, 5 % éprouvant des difficultés contrairement à l'attente. Ces derniers, plutôt jeunes, obtenaient de bonnes notes au collège, ont été reçus au diplôme national du brevet et inscrits dans des lycées réputés peu sélectifs. On peut émettre l'hypothèse que les conseils de classe ont voulu donner « leur chance » aux élèves, quitte à ce qu'ils redoublent en seconde. Les résultats de l'enquête ne confirment donc pas l'affirmation fréquente du caractère incertain du pronostic de réussite formulé par les professeurs.

L'appréciation que l'on est en droit de porter sur la pertinence du pronostic doit être nuancée à la lumière de trois observations, les deux premières concernant la méthode retenue et la troisième utilisant les résultats d'une étude conduite par la direction de l'évaluation et de la prospective (DEP) :

- l'enquête a été limitée à l'analyse du cas des élèves admis en seconde, dont les résultats n'étaient pas ceux qu'on attendait. On n'a pas cherché à savoir si des élèves proposés pour le redoublement auraient pu réussir en seconde ;
- l'enquête n'a pas pris en compte le cas des élèves qui, admis en seconde avec un pronostic défavorable, sont effectivement en situation difficile et vont redoubler. Or ils existent. Dans ce cas de figure le pronostic est de qualité, mais c'est le redoublement qui est prévu et en quelque sorte l programmé ;
- l'enquête ne s'est pas intéressée à la valeur des données sur lesquelles le pronostic est fondé. Or l'étude, publiée par la DEP en 1987, qui s'appuie sur les résultats de l'évaluation pédagogique menée en juin 1984 auprès des élèves de troisième, montre que 24 % des élèves uniformément faibles (3,6 % du total des élèves de troisième) sont admis en seconde, tandis que 20 % des élèves réussissant bien dans les six disciplines concernées par l'évaluation ou dans cinq (3 % du total des élèves de troisième), sont proposés pour le redoublement.

Les évaluations fondées sur des épreuves standardisées

L'utilisation systématique d'épreuves standardisées, où chaque « item » permet de mesurer le degré de maîtrise d'une connaissance très parcellaire, caractérise les évaluations conduites par l'administration centrale, les échelons déconcentrés ou les institutions de recherche, comme le recours presque exclusif à la notation de copies, caractérise l'évaluation pratiquée par les professeurs en classe ou lors des examens.

Les évaluations conduites par l'administration centrale

Introduites au cours des années 1970 pour l'école élémentaire, progressivement étendues au collège et approfondies, les opérations d'évaluation par grandes enquêtes, conduites par les directions pédagogiques, la direction de l'évaluation et de la prospective et les services qui l'ont précédée, visent à mettre en oeuvre une évaluation des connaissances, voulue objective, grâce à des épreuves standardisées. Chacun des exercices proposés ne peut toutefois permettre de mesurer qu'une connaissance précise, celle que cet exercice a pour seul objectif de mettre en évidence. Les acquis des élèves ne sont donc appréhendés que par juxtaposition et tous ceux qui ne sont pas mesurables par de telles épreuves ne sont pas évalués. Telle est l'une des limites des évaluations par grandes enquêtes qui ont été réalisées de 1979 à 1989.

Celles-ci répondaient au souci de mesurer l'hétérogénéité du niveau des élèves à l'entrée en sixième, alors perçue comme trop importante, et d'aider les enseignants à la prendre en compte. Elles visaient aussi à vérifier si les élèves possédaient bien les savoirs et les savoir-faire définis par les programmes et instructions et à étudier les progressions des élèves. Cet important travail – puisque l'étude relative au collège a porté sur un échantillon de 9 000 élèves de sixième dont 6 000 ont pu être suivis jusqu'en classe de troisième - a permis notamment de progresser dans la formulation d'objectifs précis et donc évaluables ; il a permis aussi d'éclairer la compréhension des programmes. La tâche ainsi accomplie a, en ce sens, contribué à améliorer, de façon certes très inégale selon les disciplines, la formulation des nouveaux programmes en la rendant plus explicite.

Cependant, l'insuffisante exploitation des matériaux de l'enquête, à vrai dire considérables, et la difficulté d'interpréter, dans une perspective globale, les résultats d'exercices correspondant chacun à des savoirs excessivement parcellisés, n'ont pas permis d'interpréter de façon claire l'information recueillie et de la mettre au service de l'institution scolaire. Il n'a pas non plus été possible de déterminer ce qui, dans les résultats obtenus, relevait de la difficulté des programmes ou de la pédagogie mise en oeuvre.

Parmi les leçons que l'on peut tirer de ces opérations, deux sont particulièrement intéressantes :

- à performances égales, deux élèves n'ont pas nécessairement été orientés de façon identique. Ainsi, des élèves qui en 1982 obtenaient en classe de cinquième les mêmes performances, pouvaient soit être admis en quatrième, soit redoubler **;**
- l'hétérogénéité des connaissances des élèves est partiellement construite par l'enseignement lui-même. On a en effet observé, d'une part, qu'il existait une relation entre l'importance qu'un professeur accorde à telle ou telle partie du programme et la maîtrise qu'ont les élèves des connaissances correspondantes et, d'autre part, que les professeurs accordent une importance variable à une même question.

Pour conduire ces enquêtes, la direction de l'évaluation et de la prospective a sollicité le concours de l'INRP[6], de L'INETOP[7] et de l'IREDU[8] en particulier, qui ont parfois participé, avec l'Inspection générale, à l'élaboration des exercices, au traitement des données et à l'analyse des résultats. Il était donc particulièrement opportun d'amorcer le recensement des apports de ces instituts, à la connaissance des acquis des élèves.

Les apports des organismes de recherche

[6] INRP : Institut national de recherche pédagogique
[7] INETOP : Institut national d'étude du travail et d'orientation professionnelle
[8] IREDU : Institut de recherche sur l'économie de l'éducation.

L'investigation ne visait pas à recenser de façon exhaustive les travaux de la recherche dans le domaine de l'évaluation des acquis, mais à mettre en évidence, à partir de quelques exemples, l'intérêt qu'il y aurait à conduire un tel recensement. L'accessibilité des travaux de la recherche conditionne, en effet, la possibilité pour les décideurs de l'éducation et pour les professeurs d'en tirer davantage parti.

Le caractère parfois sélectif des documents transmis par les trois organismes de recherche précédemment évoqués n'a pas permis de dégager les grandes orientations qui leur sont communes. On se bornera à évoquer quelques travaux de chacun d'entre eux. Ils ne seront pas nécessairement représentatifs de l'activité globale de chacun.

Une étude conduite par l'INRP à la fin des années 70 sur l'enseignement des mathématiques dans les classes de CM2 et de sixième a permis de montrer que les élèves d'aujourd'hui

- maîtrisent, aussi bien que leurs aînés, le maniement des techniques opératoires tout en utilisant efficacement des outils nouveaux comme tableaux ou diagrammes ;
- éprouvent les mêmes difficultés dans la résolution des problèmes.

Ce même institut, proposant en 1987 à des élèves la dictée d'un texte de Fénelon qui avait été donné en 1887 à des élèves d'un même niveau de scolarité, a mis en évidence que :

- les élèves d'aujourd'hui sont incontestablement meilleurs que leurs camarades d'il y a cent ans, bien que l'on consacre désormais moins de temps à l'apprentissage de l'orthographe ;
- les progrès en orthographe sont très étroitement liés au progrès dans la connaissance de la langue.

Il semblerait donc possible de considérer que la baisse de niveau, souvent invoquée, ne corresponde pas à la réalité en orthographe et en calcul, pour les enfants de moins de 14 ans.

Une telle constatation se trouve étayée par les résultats des travaux conduits sur l'enseignement des sciences, par l'INETOP, en 1975, dans le cadre d'une enquête internationale concernant une vingtaine de pays dont la France, qui ont permis de conclure que le niveau des « élites » n'était pas abaissé par une démocratisation du système scolaire.

Cet institut apporte des éléments d'information intéressants sur la valeur des pronostics de réussite scolaire, fondés sur des tests de connaissance, par rapport à ceux qui prennent appui sur les bulletins scolaires. En 1965, il montre qu'en CM2 la prédiction de la réussite des élèves en sixième est un peu meilleure si l'on utilise des tests de connaissance que si l'on se fonde sur les notes scolaires. En 1987, il met en évidence que l'appréciation globale portée sur les bulletins de la fin du premier trimestre de troisième est ce qui permet d'anticiper le mieux la progression dans le second cycle ainsi que la réussite au baccalauréat, mais que les résultats d'un « test composite d'intelligence générale », recueillis en une heure dès le début du trimestre, permettent un pronostic à long terme presque aussi bon.

L'INETOP a montré par ailleurs, dans une étude publiée en 1988, que les disciplines littéraires (français et anglais) avaient un impact plus important sur la décision d'orientation que les mathématiques dont la suprématie serait donc injustement invoquée.

L'IREDU s'est également intéressé à l'orientation des élèves mais dans la perspective d'une évaluation des procédures en vigueur en France. A partir d'un travail qu'il a conduit dans 17 collèges de l'académie de Dijon, il affirme qu'à performances égales sur des épreuves standardisées, les élèves des milieux favorisés obtiendraient de meilleures notes, et qu'à note égale ils accéderaient plus aisément à la classe supérieure. Ces travaux, que le système éducatif ne saurait ignorer, incitent à penser qu'au-delà des variables individuelles, l'avenir d'un élève est largement déterminé par l'ensemble des éléments du contexte dans lequel il est scolarisé.

<u>Les apports des travaux conduits dans les académies</u>

Quatre observations doivent être formulées.

- L'administration centrale et le monde de la recherche n'ont pas pour autant l'exclusivité des initiatives en matière d'évaluation des acquis des élèves. Il existe dans les académies, les départements, et les établissements scolaires, une grande richesse de réalisations en la matière. Le recensement a été toutefois difficile et très incomplet parce que les autorités hiérarchiques ne prennent pas encore systématiquement en compte ces sortes d'initiatives et aussi parce que leurs modalités de réalisation ainsi que l'analyse des résultats ne sont généralement pas suffisamment formalisées pour qu'on puisse en rendre compte et, a fortiori, les diffuser.
- Parmi les évaluations réalisées, celles qui concernent les acquis sont peu nombreuses. Il semblerait que la performance de l'École s'apprécie par la connaissance des flux d'élèves plus que par celle des
- Les initiatives qui ont pu être recensées reflètent l'existence d'une activité d'évaluation importante en collège (60 %), moindre à l'école élémentaire (40 %) et quasi nulle au lycée. Le cycle d'observation du collège est privilégié, ainsi qu'à un moindre degré le cours préparatoire et le cours moyen deuxième année à l'école élémentaire, c'est-à-dire les classes qui correspondent à l'entrée et à la sortie du cycle.
- La quasi-totalité des évaluations concernent le français (50 %) et les mathématiques (35 %). L'orthographe et la lecture sont particulièrement concernées.

Quelques opérations s'appuient sur des échantillons d'élèves importants, plusieurs milliers d'élèves.

L'ensemble de ces éléments montre que les opérations d'évaluation se concentrent sur les niveaux scolaires où les difficultés se révèlent avec une particulière évidence et qu'elles portent sur des compétences qu'il est possible d'appréhender grâce à des exercices écrits, simples et codifiés.

L'évaluation conçue comme un outil de formation des enseignants et de remédiation

L'action réalisée depuis **1986,** dans une académie, avec pour objectif de « former les maîtres par l'évaluation de leurs élèves », préludait à la mise en place, à la rentrée 1989**,** de l'opération d'évaluation des acquis des élèves à la fin du CE1 et du CM2. Cette opération est réellement originale en fonction de sa finalité, de son champ d'application et de ses objectifs, immédiats et à moyen terme.

Elle se justifiait par la nécessité de corriger, dès l'école élémentaire, les déficits d'apprentissage des élèves dans la mesure où l'on sait que l'échec à ce niveau interdit la poursuite d'une scolarité satisfaisante et où l'on vise à conduire 80 % des élèves de cette génération au niveau du baccalauréat. Elle conditionnait, au demeurant, la crédibilité de cette ambition**.**

C'était la première fois qu'une évaluation concernait, en France, la totalité des élèves à deux niveaux de la scolarité, soit 1 700 000 enfants.

Enfin, l'objectif immédiat recherché était de déceler les difficultés rencontrées par les élèves afin de permettre au maître de les corriger en prenant appui sur la formation mise en place en fonction de l'analyse de ces difficultés. Il ne s'agissait donc pas de tester simplement les connaissances des élèves. L'objectif à moyen terme est de ne plus considérer que les échecs constituent une fatalité, mais qu'ils peuvent être dépassés. On peut tirer de cette opération deux sortes d'enseignements : les uns concernent le fonctionnement général du système éducatif, les autres portent plus précisément sur les acquis des élèves dans deux disciplines.

La mobilisation du corps enseignant a été réelle, moins importante sans doute dans les collèges où l'opération a été perçue comme concernant davantage l'école élémentaire et où les professeurs avaient antérieurement été sollicités dans le cadre de la rénovation des collèges.

La difficulté de faire appliquer les instructions sur l'ensemble du territoire a été également observée. Même des consignes purement techniques dont le respect conditionne, à l'évidence, la validité des résultats, n'ont pas toujours été prises en compte. Il est de surcroît apparu que les interprétations successives des échelons hiérarchiques pouvaient aboutir à rendre les objectifs plus complexes et donc plus difficiles à atteindre.

Si la correction des épreuves n'a pas, en elle-même, posé de graves problèmes, il a en été différemment du codage des réponses qui a entraîné pour les enseignants une tâche fastidieuse. Mais le plus difficile a été de proposer aux maîtres des formations qui soient fondées sur les difficultés effectivement constatées et qui répondent aux besoins ressentis.

Ainsi l'opération a-t-elle non seulement permis de déceler les difficultés des élèves, mais aussi de montrer que la formation des enseignants à la remédiation est une opération complexe et que les obstacles rencontrés par les maîtres dans leurs classes sont grands. Bien que cette évaluation ait pour objectif essentiel d'aider le maître à déceler les difficultés de chaque élève, afin de pouvoir les corriger et non de déterminer le niveau de connaissances des élèves en français et en mathématiques, la direction de l'évaluation et de la prospective s'est efforcée de mettre en évidence, par agrégation des résultats, des conclusions nationales qui ont été très largement diffusées. Cinq remarques doivent être formulées afin que l'opinion n'ait pas de l'École une vision faussée.

- Il importe de distinguer les connaissances acquises à un moment donné de celles qui sont en cours d'acquisition. On pourrait, par exemple, tirer des conclusions pessimistes du fait que 45 % seulement des élèves, au tout début du CE2, savent utiliser la division. La réalité est toute différente puisque cette utilisation ne sera, d'après les programmes mêmes, exigible qu'à la fin du cours moyen.
- Les lacunes véritables ne doivent pas être confondues avec celles qui sont apparentes et temporaires, le savoir antérieurement acquis étant simplement en sommeil, faute d'avoir été récemment mobilisé, et pouvant être rappelé par de simples exercices.
- La distinction entre les lacunes dues à un manque de connaissances et celles qui résultent d'apprentissages défectueux doit également être faite.
- Les élèves semblent parfois posséder des connaissances qu'on ne leur supposait pas. Ainsi, le niveau d'élèves de zones d'éducation prioritaires, dans tel ou tel domaine, est de nature à surprendre. Il est donc très important que les opérations d'évaluation évitent que ne perdurent des représentations fausses. C'est en repérant ce que les plus faibles savent faire et ce que les plus forts font encore assez mal, que le maître aidera tous les élèves à progresser. Ainsi pourrait s'atténuer le clivage traditionnellement établi entre qualité de l'École et démocratisation.
- Le constat le plus négatif de l'enquête semble résider dans une insuffisante maîtrise des méthodes de travail. Il est important que les maîtres en prennent conscience afin d'éviter qu'à chaque niveau scolaire on ne contourne l'obstacle, soit en reportant au niveau suivant l'apprentissage des méthodes, soit en déplorant que cela n'ait pas été fait au niveau précédent.

Propositions

Les constatations dégagées, soit directement par l'Inspection générale, soit à partir d'une rapide analyse des travaux réalisés à l'initiative de l'administration centrale, des académies et des institutions de recherche, doivent conduire à la formulation de propositions. La relation forte qu'il convient d'établir, pour agir, entre l'évaluation et l'utilisation de ses résultats, a d'ailleurs été affirmée dans l'opération d'évaluation à la fin des cycles d'apprentissage et d'approfondissement qui, au-delà des constats, visait explicitement à aider les maîtres à corriger, immédiatement dans leur classe, les difficultés mises en évidence.

On s'efforcera donc de proposer des solutions ou des éléments de correction relatifs à chacun des constats importants qui ont été faits sur la notation en classe, les modalités de délivrance du diplôme national du brevet, les décisions d'orientation, l'opération d'évaluation CE2-6e et les initiatives prises par le ministère, les rectorats, les établissements scolaires ou les chercheurs, dans le domaine de l'évaluation.

Pour améliorer la notation en classe

Il ne serait pas raisonnable de continuer à utiliser 20 % du temps de la classe à la réalisation par les élèves de devoirs écrits sous la surveillance des professeurs. Ceci ne veut naturellement pas dire que les élèves doivent faire moins de devoirs et d'exercices[9], car ceux-ci constituent un moyen essentiel pour apprendre

[9] Il convient même d'accroître le nombre des travaux personnels donnés aux élèves des établissements défavorisés, afin que ces derniers aient, autant que leurs camarades des établissements favorisés, l'occasion de réaliser devoirs et exercices.

et pour progresser, mais que le temps consacré en classe à cette activité doit être réduit. Le temps d'enseignement qui aura pu être dégagé pourrait être partagé entre des activités d'aide aux élèves et de formation des enseignants.

Tous les devoirs réalisés à la maison ne sont pas notés, ce qui ne présente, en soi, pas d'inconvénients. Il importe, en revanche, que l'élève sache toujours quel est l'objectif du travail demandé exercice d'entraînement ou vérification de certaines acquisitions, et s'il en résultera ou non une note. En tout état de cause un équilibre, variable selon les disciplines et les niveaux de scolarité, entre les exercices à vocation de bilan et ceux qui ont surtout vertu d'entraînement doit être recherché. Il semble que ce ne soit pas toujours le cas et que le souci d'assurer un contrôle continu conduise à privilégier les bilans chiffrés.

Les annotations portées sur les copies, nécessaires pour aider l'élève à identifier ses progrès et ses lacunes, prennent beaucoup de temps aux professeurs. Plus de la moitié des élèves les estiment cependant peu utiles. Pour en améliorer l'efficacité il faut que les objectifs du travail aient été préalablement mieux précisés, que les connaissances et les savoir-faire qui seront évalués ainsi que les critères de jugement aient été explicités. Il convient également d'établir une progression du nombre et du niveau des exigences. Pour aider les professeurs il importera de réaliser des études dans chaque discipline sur les méthodes de correction des productions des élèves et d'organiser des stages de formation.

La note résume un grand nombre d'informations, ce qui rend son interprétation difficile, alors que les épreuves standardisées mesurent des acquisitions parcellaires qu'il faut pouvoir relier entre elles. Le professeur doit utiliser l'un ou l'autre de ces instruments d'évaluation en fonction de l'objectif visé. C'est en les associant et en recoupant leurs résultats qu'il parviendra à mieux évaluer les performances de l'élève. Il appartient à l'administration centrale de mettre à sa disposition des banques d'épreuves.

Pour améliorer les modalités de délivrance du diplôme national du brevet et son utilité

Les taux de réussite au diplôme national du brevet ne varient pas seulement en fonction du niveau de connaissance des élèves, mais également, en proportion non négligeable, avec le degré de sévérité de la notation en quatrième et troisième, prise en compte pour le contrôle continu, et avec le taux de repêchage pratiqué dans le département.

Pour améliorer la rigueur du contrôle continu, il convient de faire un effort similaire à celui qui a été réalisé pour les épreuves écrites, de proposer, au niveau national, dans chaque discipline, des références précises sur ce que l'on doit exiger des élèves et de réduire le recours à la procédure de repêchage. Le niveau d'exigence doit être fixé à la fois en fonction des performances actuelles des élèves et des ambitions du système éducatif. Il serait en effet très dangereux que le taux de réussite socialement acceptable au diplôme national du brevet constitue un facteur important de régulation du niveau requis pour la certification des études à la fin du collège.

Ce diplôme, qui mobilise du temps et des moyens financiers, n'est pas toujours considéré comme très utile, bien qu'il ait valeur d'entraînement aux examens et concours ultérieurs, qu'il soit pour les élèves un facteur d'incitation au travail, qu'il constitue - c'est très important pour certains parents - la première reconnaissance officielle d'un niveau d'étude et que le contrôle continu soit une occasion supplémentaire de concertation entre les professeurs. Le diplôme national du brevet n'aura sa pleine efficacité que dans la mesure où, conformément aux objectifs initiaux, il constituera un véritable instrument d'évaluation qui permettra :

- au collège d'apprécier les résultats obtenus en les comparant à ceux d'autres établissements, de les analyser pour repérer les difficultés des élèves et d'améliorer la qualité des enseignements ;
- aux professeurs de seconde des lycées d'accueil, de prévoir des remises à niveau en fonction des lacunes identifiées et d'organiser les progressions en meilleure connaissance de cause.

Pour améliorer l'efficacité des opérations d'évaluation nationales ou académiques

L'opération « CE2-6e » a montré que la mobilisation des enseignants était grande quand les objectifs de l'évaluation étaient clairs et quand l'utilité immédiate de la démarche pour corriger les difficultés constatées chez les élèves était perçue. Il est donc souhaitable que toute évaluation, dont l'initiative émane de l'administration centrale ou des échelons académiques soit précédée d'une information précise et complète qui intègre la façon dont chaque enseignant pourra en tirer parti, immédiatement ou à terme.

Constater les difficultés ou les lacunes des élèves ne saurait être positif pour les enseignants s'ils n'étaient mis en situation de les corriger. Or la difficulté la plus grande est justement dans le passage du constat au remède. Il faut donc que dans toute la mesure du possible les évaluations débouchent sur des recommandations et propositions de correctifs et sur une aide aux enseignants. Ces derniers doivent pouvoir accéder à l'analyse des résultats des évaluations et des causes des difficultés constatées. Ils doivent pouvoir recourir à des outils informatiques qui leur permettent de traiter les données relatives à chacune de leur classe, voire à chacun de leurs élèves, et d'être préparés à utiliser des instruments de remédiation ou à les construire.

L'utilisation des évaluations, en particulier par les maîtres eux-mêmes, doit devenir une préoccupation permanente des académies. Ceci exigera un effort important de formalisation des travaux, d'analyse et de diffusion des résultats, ainsi que de conception et de mise en oeuvre de formations adaptées.

Pour améliorer l'utilité des travaux de recherche

Les travaux des instituts de recherche prennent rarement les acquis des élèves comme objet d'étude en tant que tel, les résultats en sont peu connus et quelquefois peu aisément accessibles. Certains d'entre eux débouchent cependant sur des constats que nul ne doit ignorer.

Il importe donc d'en effectuer un recensement complet, en organisant la collaboration entre ces organismes; la direction de l'évaluation et de la prospective du ministère, les directions pédagogiques et les corps d'inspection. La diffusion des résultats concernant les acquis des élèves aux différentes étapes de la scolarité et les niveaux d'exigence de connaissance requis pour le passage d'un cycle à l'autre, pourrait être réalisée par le Centre national de documentation pédagogique, en liaison avec la direction de l'évaluation et de la prospective.

L'administration centrale et les académies devront rechercher des solutions aux problèmes qui auront été identifiés, soit en conduisant elles-mêmes des études ou des expériences, soit en lançant des appels d'offres. Elles pourront solliciter davantage les chercheurs pour la réalisation de travaux sur l'analyse des difficultés et la recherche de remèdes appropriés**.**

Mais il est peut-être plus important encore de tirer parti des évaluations déjà effectuées que d'en prévoir constamment de nouvelles, sauf dans certaines disciplines jusqu'à présent peu étudiées de ce point de vue.

Pour améliorer l'orientation des élèves

La cohérence observée, dans les onze collèges étudiés, entre les résultats des élèves en troisième et en seconde incite à conclure que le pronostic formulé par les professeurs est bon. Cependant le taux de redoublement en seconde est élevé. Avec des résultats identiques à des épreuves standardisées, deux élèves n'obtiennent pas toujours la même note et, avec la même note, n'accèdent pas nécessairement au même type de scolarité. Les conseils de classe ne consacrent de surcroît à la discussion sur le devenir de l'élève que moins du quart des interventions. Il semble donc possible d'améliorer la qualité de l'orientation.

Le redoublement en seconde pose une question particulière dans la mesure où il semble prévu par les professeurs de collège, qui souhaitent donner leur chance à un grand nombre d'élèves. Des dispositions d'ordre pédagogique devraient permettre de ne pas renoncer à cette ambition. Deux voies complémentaires s'offrent dans cette perspective : mieux préparer les élèves à l'entrée en seconde, prévoir leur mise à niveau éventuelle dès l'entrée au lycée.

Puisque les épreuves standardisées et les notes apportent des informations différentes sur l'élève, il importe de combiner l'usage de ces deux modes d'évaluation. Pour parvenir à mieux fonder les décisions d'orientation, il faut également que les conseils de classe utilisent toutes les sources d'information disponibles concernant, d'une part les performances de l'établissement, mesurées par comparaison avec celles de collèges de même type, et d'autre part, l'évolution scolaire de l'élève et ses projets. Ceci exige une meilleure organisation du partage de l'information par l'ensemble des membres du conseil - en recourant en particulier à l'informatique - afin de consacrer l'essentiel du temps du conseil à l'analyse en commun de la situation de chaque élève - en suscitant notamment la contribution des délégués - et au choix de la meilleure voie d'orientation.

La prise en compte du projet de l'élève et des avis des délégués ne doit cependant pas conduire l'établissement à abdiquer une part de ses responsabilités en renonçant à apporter à l'élève des indications précises et argumentées sur ses chances de réussite.

Conclusion

L'évaluation des acquis des élèves n'est évidemment pas une fin en soi. Elle ne vaut que comme instrument de régulation du système scolaire à tous les niveaux et d'aide à la décision.

Les constats de dysfonctionnement ont permis de formuler des propositions pour améliorer la qualité des évaluations. Ces propositions peuvent être regroupées selon six grandes directions d'action.

Améliorer la précision des objectifs de l'évaluation

L'objectif de chaque évaluation des acquis des élèves, en classe, aux examens et dans le cadre de grandes enquêtes, devra toujours être clairement précisé afin de permettre :

- aux élèves de mieux tirer parti des notes et appréciations des professeurs ou des résultats des épreuves standardisées ;
- aux enseignants d'utiliser les résultats de toutes les évaluations pour ajuster leur enseignement ;
- aux décideurs de prendre appui sur les conclusions des évaluations pour prendre les décisions.

Une définition plus précise des objectifs de l'enseignement, dans chacune des disciplines et à chacun des niveaux scolaires (cycle ou classe) conditionne la possibilité de mesurer efficacement le degré de maîtrise par les élèves des connaissances et des savoir-faire attendus. Mais un excès de précision mettrait en cause l'adaptation de l'enseignement dispensé au niveau effectif des élèves, à tel ou tel moment, et donc la qualité même de l'enseignement. Le professeur enfermé dans un cadre trop strict ne disposerait plus d'une suffisante liberté intellectuelle et se transformerait en simple répétiteur. Un juste équilibre doit être recherché.

Accroître le nombre d'instruments et de méthodes d'évaluation et les améliorer

Le décalage perçu entre l'abondance du discours sur l'évaluation et la réalité des pratiques est dû au fait que les instruments susceptibles d'être utilisés sont encore peu nombreux et que ceux qui existent sont insuffisamment connus.

Il conviendra donc de recenser les instruments existants et d'en assurer largement la diffusion. Il serait utile de mettre en place, aux niveaux national et rectoral, des banques d'épreuves accessibles aux enseignants. La construction de nouvelles épreuves et la mise au point de méthodes mieux adaptées de correction des copies devraient être réalisées en associant des équipes d'enseignants conscients de leurs besoins pédagogiques, des spécialistes de l'évaluation et des chercheurs.

Mieux diffuser l'information

L'évaluation des acquis des élèves nécessite une diffusion de l'information sur :

- les moyens à mettre en oeuvre (consignes de passation d'épreuves standardisées, critères de correction, barème de notation...) ;
- l'interprétation des résultats obtenus (signification d'une note attribuée à un devoir, dans une discipline et à un moment précis de l'année, conclusion à tirer de l'échec à l'un des exercices d'une épreuve standardisée...) ;
- les facteurs à prendre en compte pour la décision, qu'il s'agisse de certifier une formation (importance relative de l'écrit et du contrôle continu au diplôme national du brevet) ou d'orienter un élève (âge, résultats scolaires, projet personnel...) ;
- les éléments qui permettent ou non de fonder un jugement (le fait que tous les élèves ne sachent pas effectuer une division en CE2 ne signifie pas que l'école n'apprend plus à compter).

Or il apparaît, par exemple, dans les conseils de classe ou dans la diffusion des résultats du brevet que le partage de l'information est actuellement insuffisant. Il faut donc lever les obstacles qui procéderaient d'une conception trop hiérarchique des relations à l'intérieur du système éducatif, de l'habitude du secret ou de l'utilisation de moyens de communication inadéquats.

Valoriser le temps que le professeur consacre à l'évaluation

La surveillance de devoirs réalisés en classe, qui occupe près du cinquième du temps d'enseignement en troisième et en seconde, n'est pas une activité d'évaluation. En revanche, l'interrogation écrite de quelques minutes ou l'interrogation orale en est une. Des indications doivent donc être données aux enseignants pour que le partage du temps entre l'enseignement, l'aide personnelle aux élèves et l'évaluation soit mieux équilibré. La correction des copies nécessite un temps important, qu'il paraît difficile de réduire, mais qui peut être mieux rentabilisé, si notes et appréciations permettent d'éclairer plus efficacement les élèves. Des aides méthodologiques doivent être proposées aux professeurs.

Le temps consacré à l'échange d'informations lors des conseils de classe doit être réduit au profit de l'analyse par les professeurs des résultats des élèves pour préparer des décisions pédagogiques ou d'orientation.

Mieux utiliser l'évaluation pour améliorer le fonctionnement du système éducatif

Les évaluations des acquis conduites par l'administration centrale, par les académies ainsi que par les organismes de recherche n’ont pas eu, jusqu'à présent, une grande influence sur la prise de décision. Il pourrait en aller différemment :

- si l'analyse des opérations réalisées était approfondie et si les résultats étaient plus aisément accessibles ;
- si le nombre des nouvelles évaluations était limité et si leurs objectifs étaient définis avec plus de précision, comme ce fut le cas pour l'opération d'évaluation CE-6^{e}.

Améliorer la part des analyses qualitatives dans la gestion du système éducatif

L'objectif qui vise à conduire 80 % des élèves au niveau du baccalauréat a pu inciter l'institution à privilégier les indicateurs quantitatifs, et les décideurs à infléchir le fonctionnement du système pour que ces indicateurs varient dans le sens positif. Ainsi les taux de redoublement ont-ils baissé et les taux de passage en seconde ont-ils augmenté sans qu'on se fonde véritablement sur l'évaluation des connaissances des élèves. Ce constat n'a encore rien d'inquiétant dans la mesure où il convenait de combattre des attitudes trop sélectives. Il ne serait cependant pas raisonnable de procéder ainsi indéfiniment.

Le recours à des critères quantitatifs n'est pas exclusif et le système scolaire met en oeuvre, parallèlement, des procédures d'évaluation à visée qualitative. C'est en aidant les maîtres à identifier les difficultés que l'on a cherché, dans l'évaluation CE2, à améliorer la performance des élèves et celle de l'école élémentaire.

Il n'en demeure pas moins que la volonté de rigueur dans les méthodes d'évaluation peut s'opposer à des exigences d'un autre ordre, par exemple le taux d'échec socialement acceptable au diplôme national du brevet.

Pour assurer le « pilotage » du système éducatif, il conviendra de rechercher un meilleur équilibre entre le recours à des critères quantitatifs et l'utilisation d'évaluations qualitatives.

ANNEXE 6 : EVALUATION INDIVIDUELLE ET EVALUATION-BILAN, DES CONTRAINTES DIFFERENTES

Les considérations qui suivent tendent à étayer l'idée, exposée dans le texte du rapport, selon laquelle il est important de clarifier, dès la phase initiale de la construction d'un dispositif, l'usage qui sera fait des données à recueillir.

Echantillonnage des items

L'un des premiers problèmes à résoudre lors de la conception d'un instrument d'évaluation est celui de l'échantillonnage des variables (i.e. des acquisitions à évaluer) :

- dans la perspective d'évaluation individuelle, on aura besoin d'un instrument qui présente des qualités spécifiques. La base de référence étant dans ce cas, plutôt les démarches d'apprentissage et les processus transférables, sans nécessité de couvrir l'intégralité d'un programme, il faudra éprouver la pertinence des hypothèses sur les hiérarchies de fonctionnement intellectuel des sujets, et la qualité de l'épreuve qu'il faudra contrôler d'abord sera la validité de construction. Si l'on veut par ailleurs pouvoir formuler un pronostic d'adaptation à partir des résultats, c'est alors à la validité empirique (ou pronostique) qu'il faudra s'intéresser. Notons que la préparation des questions d'un examen conduisant à un diplôme, et plus encore, d'un concours d'admission devra tenir compte de cet aspect (en effet, si le diplôme apparaît comme l'attestation d'une formation suivie avec succès, il doit aussi être le garant d'une bonne adaptation ultérieure à un nouveau cycle d'étude, ou à la vie professionnelle, ou à celle de citoyen).

- dans la perspective d'évaluation du système, le problème est formellement plus simple : il s'agit en principe de construire une épreuve qui couvre le champ des différents apprentissages compris dans le programme. Le type de validité sur lequel on fera porter l'accent est celui de la validité de contenu. Mais dans ce cas, la recherche d'une meilleure couverture du programme peut conduire à construire un grand nombre d'items, qu'il faudra alors répartir en sous-épreuves distinctes, administrées à des sous-échantillons distincts de sujets, en même temps qu'une sous-épreuve commune, d'ancrage, permettant l'*equating*. Il conviendra alors que les épreuves satisfassent à de nouvelles contraintes, pas toujours nécessaires dans la première perspective : indépendance locale entre items, invariance des paramètres d'item, unidimensionnalité éventuelle de l'épreuve totale (cas du modèle MRI de base).

Par ailleurs, le souci de couvrir de façon exhaustive le programme officiel conduit à s'interroger sur le niveau d'adéquation entre le programme prescrit et le programme tel qu'il est enseigné dans la classe. Ce problème comporte deux aspects : d'une part celui du dépassement du programme intentionnel par des apprentissages effectifs (d'attitudes, mais aussi de démarches intellectuelles) qui ne font pas explicitement partie du référentiel officiel (et dans ce cadre il peut sembler intéressant, dans une phase préparatoire à l'évaluation, d'explorer ce domaine) ; d'autre part, celui de l'incomplète réalisation du programme par l'enseignant au terme de l'année scolaire (ou du cursus) : en termes de l'explication du fonctionnement du système, peut-on attribuer la même signification à des lacunes observées chez les élèves selon qu'ils ont ou non été " exposés " à l'unité d'enseignement correspondante ? Il paraît évident qu'il n'en est rien.

C'est pourquoi, par exemple, dès 1970, l'IEA introduisait dans son dispositif international d'évaluation, un questionnement des professeurs sur l'« *opportunity to learn* » (l'occasion effective d'apprendre) pour chaque question posée dans ses instruments. On retrouve un souci voisin, dès l'enquête du MEN de 1980-84 sur les apprentissages au collège, dont les résultats publiés permettaient de comparer, pour chaque item le niveau de la population et le pronostic de réussite formulé par les enseignants. Toutefois ces couples de résultats se prêtaient mal à une interprétation, puisque le pronostic des maîtres était conditionné non seulement par l'occasion d'apprendre, mais aussi par la difficulté d'un tel apprentissage pour les élèves considérés. Il reste que ce type d'information nous semble avoir tout à fait sa place dans une évaluation de type bilan, malgré la lourdeur nécessaire de la procédure, alors que son utilité dans une évaluation de type diagnostic paraît beaucoup plus discutable.

Niveau de difficulté des items

Le problème du choix de la difficulté des items peut sembler à première vue un faux problème. A priori, on peut penser en effet que, puisqu'il s'agit d'établir un bilan, il n'y a pas vraiment lieu de s'en préoccuper : on s'en tient à construire des questions qui correspondent au programme, puisque l'objectif est justement de contrôler dans quelle mesure ce programme a été " acquis ". Cependant quand on passe à la réalisation, on peut constater que les choses ne sont pas si simples. Une première difficulté tient au fait que, même si ces programmes ont été formulés en termes d'objectifs pédagogiques, ces formulations ne sont pas opérationnelles, et laissent varier plus ou moins largement les caractéristiques de la performance attendue. Ceci n'est d'ailleurs probablement pas regrettable : il est sans doute bon que le maître dans sa classe puisse adapter le niveau de difficulté de son enseignement aux capacités d'assimilation de ces élèves. Un même objectif peut donc souvent faire l'objet de différents items d'évaluation, de difficulté gradée.

Par ailleurs, dans de nombreux cas, la difficulté objective d'un item sera affectée par tout un ensemble de caractéristiques étrangères à la dimension qu'on veut mesurer (ce qui fait qu'un item isolé est en fait peu fidèle). L'énoncé de la question lui-même intervient, à la fois par sa forme (choix des mots, longueur des phrases…) et par son contenu (e.g. informations en surnombre dans un énoncé de problème). Le " format " de l'item joue également un rôle, par ses différents aspects : le mode de questionnement (QCM, question ouverte, à complètement, …), le matériel (verbal, figuratif, …), le mode de réponse, le nombre d'éventualités (cas d'un QCM), etc… Le positionnement de l'item dans l'épreuve (au début, à la fin, …) va également moduler la difficulté.

On peut donc agir et modifier ces différents aspects pour moduler la difficulté des items. On aura intérêt, dans le cadre de la construction d'un instrument en vue d'un bilan, à obtenir majoritairement des items de difficulté intermédiaire, qui discriminent mieux entre les sujets.

Ces préoccupations cesseront d'occuper le premier plan si l'instrument est construit en vue d'un diagnostic. Dans ce dernier cas, en effet, on ne cherche pas à ajuster le niveau de difficulté, ni à obtenir une distribution générale des performances proche de la distribution de Gauss, mais plutôt à construire des items qui permettent de bien identifier la démarche de résolution du sujet, et c'est plutôt sur les différentes éventualités de réponse pour chaque question qu'on se centrera.

Codage des réponses

Dans la perspective diagnostique, on voit l'intérêt de prévoir une correction analytique, basée sur l'existence de différents types de démarches, erronées ou non, conduisant au choix d'une éventualité de réponse par le sujet. Encore faudrait-il que chaque type de démarche ainsi identifiable puisse être re-situé dans le cadre d'un modèle psychopédagogique spécifié. Il faudrait à tout le moins contrôler pour chaque item, la liaison entre chaque éventualité de réponse et le niveau général des sujets qui l'auront choisie.

Dans l'évaluation-bilan, par contre, les distinctions entre types de mauvaises réponses, et la proposition de quasi-bonne réponse (codée 2), ne font que risquer d'introduire des "bruits" parasites dans le relevé des informations.

Quantitatif / qualitatif.

Au terme d'une évaluation de type bilan, on doit pouvoir disposer d'une information quantifiée. Cette quantification comprend la construction de scores globaux et de sous-scores pour les différentes dimensions éventuelles d'une épreuve (e.g. pour les mathématiques : géométrie, opérations numériques, raisonnement mathématique…), ceci d'une part pour rendre compte au public, d'autre part pour permettre les mises en relation, et comparaisons entre sous-groupes, ou dans le temps ou l'espace, en utilisant l'outil statistique. Cependant, pour pouvoir ajouter des scores obtenus à différents items, il est nécessaire d'éprouver l'unidimensionnalité (ou d'identifier les différentes dimensions) sous-jacente(s) à une épreuve d'évaluation. Les traitements statistiques correspondants pourront être entrepris dès le stade de la pré-expérimentation.

Dans une évaluation de type diagnostic au contraire, on est plutôt axé sur les informations qualitatives apportées par l'évaluation. On s'y intéresse au degré d'atteinte du critère par chaque sujet plutôt qu'à son positionnement par rapport à la norme du groupe. On étudie séparément les réponses à chaque groupe d'items représentatif d'une progression de la démarche intellectuelle vers un objectif local. Dans ce contexte, l'uni- ou la pluri-dimensionnalité globale de l'épreuve ne présente qu'un caractère accessoire, et ne constituent pas des conditions *sine qua non* de l'utilisation qu'on fera des résultats.

ANNEXE 7 : ANALYSE DES PROTOCOLES D'HISTOIRE-GEOGRAPHIE :

Nous avons procédé à une analyse concernant l'histoire-géographie à partir des évaluations de 1984 et 1995. Nous présentons ici la synthèse de ce travail.

1. Portent-elles sur l'ensemble des connaissances acquises par l'élève ou sont-elles réduites aux connaissances acquises en troisième (représentativité des exercices par rapport aux instructions officielles) ?

Ces évaluations portent sur « *la totalité de la scolarité au collège »* (1984, document enseignant, p.2). Pour 1984, le choix des exercices repose sur les textes officiels de 1977 (document de travail, 1984, p.2) : une analyse de ces textes a permis l'élaboration d'un tableau d'objectifs (p.6 et 7). Pour 1995, l'évaluation repose sur les textes de 1985 et 1987 (document enseignant, p.7) avec la conservation d'un grand nombre d'exercices de 1984 afin de pouvoir effectuer des comparaisons. Un tableau d'objectifs est également fourni (p.8 et 9) mais l'intitulé des objectifs est parfois différent de celui utilisé en 1984. Ainsi en 1984 on distinguait 4 types d'objectifs généraux (connaissance, méthodes, savoir-faire, attitudes) et seulement 2 types en 1995 (connaissances et méthode).

Pour 1984, chaque objectif général se décompose en objectifs spécifiques qui sont ensuite opérationnalisés en exercices. Pour 1995, la notion d'objectif spécifique n'est plus utilisée. Le plus souvent, à chaque objectif identifié correspond un exercice. L'analyse de cette première étape dans l'élaboration des épreuves nous questionne sur les points suivants :

- La terminologie employée dans les tableaux ne fait pas explicitement référence à un contenu. L'évaluation semble alors porter sur chaque objectif défini, mais ces objectifs sont assez larges (exemples : "*Connaître quelques grands hommes*", "*Posséder quelques grands repères*") et expriment plutôt des capacités assez générales que des connaissances précises.

- Les objectifs identifiés représentent-ils bien l'ensemble des objectifs figurant dans les instructions officielles (pour les 4 années du collège) ou seulement une partie de ces objectifs ? Il est difficile de répondre à cette question car les programmes que nous avons consultés (ceux de 1985) sont rédigés en termes de contenus thématiques (exemples : "*L'espace européen*", "*La France entre les deux guerres mondiales*") et non pas en termes d'objectifs. Mais une analyse rapide de ces instructions montre que certains aspects des programmes ne figurent pas dans ces évaluations (ainsi pour l'histoire, par exemple très peu d'items (voire aucun) portent sur le moyen âge par rapport aux nombreuses questions sur le XXéme siècle). Une analyse plus fine pourrait être menée sur ce point afin de vérifier si l'épreuve est bien représentative de l'ensemble du programme, et des quatre niveaux de classe (notion d'échantillonnage des items). A travers les documents analysés cette représentativité ne semble pas assurée de façon satisfaisante.

- Aucune indication n'est fournie sur la correspondance des exercices avec le niveau scolaire dans lequel ces connaissances doivent être transmises. Ce qui, d'une part, ne permet pas de vérifier que les évaluations portent bien sur des connaissances acquises dans les différents niveaux scolaires ($6^{ème}$ à $3^{ème}$) et, d'autre part, dans l'interprétation des résultats, ne permettra pas de repérer à quelle étape de la scolarité relier les échecs éventuels des élèves.

2 Comment ces évaluations sont-elles constituées (élaboration des exercices, type d'exercices, types de réponses, fiabilité de la mesure…)? Comment sont-elles interprétées ?

Un groupe de travail a élaboré les exercices à partir des grilles d'objectifs. Chaque exercice évalue un objectif à travers plusieurs questions. L'évaluation de 1984 évalue ainsi 42 objectifs à l'aide de 36

exercices[10], celle de 1995 comporte 32 objectifs et 33 exercices. Chaque exercice comporte plusieurs questions et peut être accompagné de supports divers : carte à compléter, photo, texte à lire…Les questions étant le plus souvent « fermées » (QCM…), la réponse de l'élève (item[11]) est souvent réduite à placer une croix dans la bonne case.

Même si effectivement ce type de questionnement facilite le traitement statistique des réponses, il présente l'inconvénient de réduire considérablement l'éventail des capacités évaluées (par exemple les capacités d'organisation des connaissances, de logique dans la réflexion, de verbalisation …).

Quelle est la fiabilité de la mesure ? Deux remarques sur ce point :

1) Lorsqu'un objectif (et de plus, comme ici, quand il est formulé de manière très générale[12]) n'est évalué que par un seul exercice, voire dans certains cas par seulement une ou deux questions, l'évaluation reste dépendante (au moins en partie) de la situation spécifique de l'exercice (contenu, présentation, support associé, type de réponse…). Cette évaluation n'évalue donc qu'une partie seulement (une des facettes) de l'ensemble de l'objectif considéré[13]. Pour pouvoir évaluer de manière fiable de tels objectifs, l'évaluation doit reposer sur plusieurs exercices de façon à varier les contenus, les supports, le type de questionnement… C'est dans ces conditions que la généralisation de la performance peut être effectuée. Par contre si l'évaluation porte clairement sur un contenu précis (délimité) il est plus facile alors de déterminer, à partir d'un nombre plus limité de situations, si l'élève possède les connaissances relatives à cet ensemble.

2) On peut observer parfois des écarts entre ce qui est évalué réellement dans un exercice et ce qui *censé* être évalué (lien ici avec la notion de validité/fiabilité de la mesure). Par exemple, si l'on se réfère au tableau général d'objectifs, l'exercice 1-5 de 1995 évalue l'objectif «*Connaître la notion de développement »* (qui est l'un des objectifs généraux de connaissances du domaine économique) alors qu'il s'agit en réalité, lorsqu'on procède à une analyse de la tâche proposée à l'élève, non pas d'évaluer des connaissances mais d'évaluer des capacités en compréhension de lecture ! Ainsi, un élève maîtrisant correctement la lecture peut réussir cet exercice en n'ayant aucune connaissance dans le domaine évalué. Et lors de l'interprétation des scores celle-ci reposera sur la maîtrise de l'objectif censé être évalué et non pas sur le contenu réel de l'évaluation… Dans ces conditions la fiabilité de la mesure n'est pas assurée. Même si ces cas doivent être rares, il faut rappeler qu'une analyse rigoureuse de chaque exercice est indispensable pour garantir la fiabilité de l'ensemble de l'épreuve (notion de validité).

Pour 1984 une analyse détaillée des résultats de chaque exercice (objectif spécifique) est fournie mais aucun résultat n'est calculé pour chaque objectif général. Pour 1995, l'analyse des résultats par exercice est beaucoup moins détaillée et une moyenne de réussite pour chaque objectif général est calculée.

On retrouve dans les commentaires des résultats de 1995, la tendance (déjà évoquée) à généraliser les résultats d'un exercice particulier à une « capacité générale » exprimée ici en terme d'objectif "large" (exemple : « *posséder des trames de repérages », « savoir dater un événement à partir de certains indices »*) et non à interpréter ces résultats en terme de connaissances acquises et/ou en terme de contenu précis[14].

Pour 1995, une comparaison des scores de réussite entre objectifs généraux est réalisée (p.59) mais il est délicat d'interpréter des différences de scores si, au préalable, on ne s'est pas assuré que ces "sous épreuves" sont bien du même degré de difficulté (observer ainsi que "*les items de connaissance en*

[10] Certains exercices sont composés de plusieurs parties, chaque partie pouvant évaluer un objectif différent, ce qui explique ce décalage.
[11] Pour les services de la DPD, l'item représente la réponse de l'élève alors que, traditionnellement, en psychométrie, l'item représente plutôt l'ensemble « exercice et réponse » car à chaque situation (exercice) ne correspond qu'une seule question, donc qu'une seule réponse. Ce qui n'est pas le cas ici, un exercice contient plusieurs questions donc plusieurs items.
[12] Par exemple : « Savoir identifier une époque ou un régime politique à partir d'indices » (objectif 1.4 de 1995), « Savoir repérer des ruptures et des phases » (objectif 2.4 BC de 1995)…
[13] Voir aussi sur ce point nos conclusions dans les rapports sur les banques d'exercices de la DPD
[14] On retrouve ici les mêmes remarques que nous avons déjà formulées à propos de l'évaluation de « compétences générales » dans les évaluations nationales de la DPD.

géographie (domaine spatial) sont bien mieux réussis (65%) que ceux du domaine chronologique (45%)" peut amener le lecteur à retenir que les élèves sont meilleurs en géographie qu'en histoire[15] si on ne lui précise pas que cette réussite supérieure pourrait également s'expliquer tout simplement par des exercices plus faciles en géographie et non pas par un niveau de connaissance supérieur).

3 Quelles comparaisons temporelles sur l'évaluation des connaissances des élèves sont réalisées à partir de ces évaluations ?

Dans cet objectif de comparaison, l'évaluation de 1995 reprend un certain nombre d'exercices de 1984. Ainsi 18 exercices[16] sont communs (au moins en partie) aux deux évaluations. Une comparaison des résultats a donc été réalisée par les services de la DPD (Les dossiers, p.59). Mais pour effectuer une analyse comparative il ne suffit pas que les exercices soient communs et strictement identiques, il est également nécessaire que les conditions de passation et de cotation le soient aussi. Or, en 1984, les élèves avaient un temps limité pour chaque exercice pour le cahier A alors que, pour les autres cahiers de 1984 et pour l'évaluation de 1995 l'élève disposait d'un temps limité pour l'ensemble du cahier (gestion libre de son temps à l'intérieur des 55 minutes). Ainsi 5 exercices communs sont issus du cahier A et dans 4 cas sur 5, les scores de 1984 sont inférieurs aux scores de 1995. L'influence de cette différence dans les conditions de passation n'est pas évoquée dans les interprétations des résultats. Pourtant cette hypothèse n'est pas négligeable et pourrait expliquer, au moins en partie, l'augmentation observée entre 1984 et 1995.

Pour effectuer une comparaison à partir d'items communs, et passés dans les mêmes conditions, il faut aussi s'assurer que ces items possèdent les qualités psychométriques requises. En particulier ils ne doivent pas présenter de biais. Une illustration de biais possible nous est donnée par l'exercice 1.7 pour lequel les auteurs notent, entre 84 et 95, une baisse des connaissances de dates historiques, sauf pour l'item portant sur l'année de la découverte du continent américain. Or, en 1992 un film est sorti en salle sur cette époque, avec un grand acteur français dans le rôle principal, avec comme titre "1492". Cette connaissance peut donc être expliquée par un fait d'actualité et il est possible qu'en 2001, si l'on proposait cet item à des élèves de 3ème, on retrouve un score à cet item plus proche de celui de 1984 que de celui de 1995. Cet item n'est donc pas un "bon "item pour pouvoir effectuer de manière fiable une comparaison des connaissances. De manière plus générale la recherche d'éventuels biais entre les deux évaluations ne semble pas avoir été effectuée. Une telle analyse doit être pourtant menée lors de telles comparaisons avec éventuellement comme conséquences une sélection des items (on retire alors de l'analyse comparative les items présentant des biais trop importants).

Une autre remarque concerne les interprétations des variations de scores entre 1984 et 1995 qui ne reposent pas sur des indicateurs statistiques, comme l'est, par exemple, l'indication du caractère *significatif*[17] d'une différence entre deux scores. Ainsi, par exemple (p.59) *"les objectifs de méthode sont, eux aussi, mieux maîtrisés en 1995 (76 %) qu'en 1984 (73%)*" semble être une interprétation un peu rapide car seulement 3% de différence entre les deux scores et aucune indication sur le caractère significatif de cette différence. Dans la même logique on parle de "*régression"* des résultats pour l'exercice 2.12 pour un passage de 44% à 40% de réussite (p.61), de "*baisse des connaissances*" pour le domaine social et culturel pour un passage de 55 à 50% (p.63)…

Des indications en terme de différence significative, mais aussi en terme d'importance de cette différence[18], seraient pertinentes pour pouvoir interpréter de façon fiable les différences de réussite observées entre les deux évaluations.

En l'absence de telles indications une relative prudence pourrait accompagner les interprétations des

[15] Ce qui n'est pas dit explicitement dans le commentaire, mais qui peut être interprété comme tel par le lecteur.
[16] 103 items communs.
[17] Toute mesure présentant une erreur de mesure (que l'on peut estimer), une différence entre 2 scores doit dépasser une certaine valeur pour être considérée comme significativement différente de 0. Ainsi l'interprétation d'une petite différence, si elle n'est pas significative, n'a pas réellement de sens.
[18] Rappel : une petite différence entre 2 scores peut être significative (lorsque, par exemple, l'échantillon de sujets est très élevé, comme dans le cas de ces évaluations) sans obligatoirement être importante et correspondre réellement à un niveau de performance (ou de connaissance) plus élevé.

différences de faible ampleur.

Un conseil de prudence sur ce point apparaît pourtant, mais situé dans la seconde partie du document[19] (page 166) : "*En définitive, ceci nous conduit à considérer comme véritablement significatives les évolutions d'amplitude supérieure à cinq points. En dessous, les résultats ne peuvent être commentés qu'en terme de tendance"*. Pourquoi les auteurs n'ont-ils pas suivi leurs recommandations dans leurs propres commentaires ?

De façon anecdotique, mais significative du type d'interprétation donnée aux résultats, on note, en 4ème page de couverture (Les dossiers), dans un court texte qui résume les résultats de 1995 : "*Par rapport aux deux évaluations antérieures (1984 et 1990), on note(…)une progression assez marquée*[20] *en mathématiques, en sciences de la vie et de la terre et en histoire-géographie*[21]*."*

Effectivement sur les 103 items communs le score de réussite passe de 50% à 55% (pas d'évaluation en 1990) mais compte tenu de nos réflexions précédentes une telle affirmation mériterait d'être nuancée.

[19] Cette seconde partie "approfondissements", située juste avant les annexes, est sans doute moins lue par les enseignants. Pourtant un certain nombre d'indications pertinentes (sur les limites de l'interprétation des scores de réussite par exemple) mériteraient de figurer en début de document.
[20] Souligné par nous.
[21] Souligné par nous.

ANNEXE 8 : ANALYSE DES PROTOCOLES DE PHYSIQUE-CHIMIE

En Sciences physiques, les évaluations nationales de la DP&D/DEP ont eu lieu en fin de 3° en 1984, 1995 et 1999 (en 1990, les Sciences physiques n'ont pas été évaluées). Nous analyserons ces évaluations en précisant leur nature et en nous interrogeant sur leur représentativité et leur qualité. Enfin, nous aborderons le problème de la comparabilité des mesures à travers le temps.

1. Que mesurent les protocoles ?

Les évaluations portent-elles sur l'ensemble des connaissances acquises par l'élève à l'issue du cycle d'orientation ou sont-elles réduites aux connaissances acquises en troisième ?

En 1984, il est affirmé que les aspects disciplinaires de l'évaluation *"consistent à apprécier auprès des élèves la maîtrise des savoirs, savoir-faire, méthodes et attitudes spécifiques à ce cycle"* (d'orientation) et "*le maintien* ***d'acquisitions antérieures au cycle d'orientation***". En 1995, on précise *que "les exercices mesurent ce que les élèves sont capables de faire (savoirs, savoir-faire, maintien des acquisitions antérieures) au terme de la classe de troisième ".* En 1999, le même type de formulation est repris dans le "Document à l'attention du professeur".

Nous analyserons la correspondance entre les objectifs d'évaluation et les objectifs des programmes pour l'année 1984, seule année pour laquelle la méthode d'élaboration est explicitement détaillée dans le Document de travail 1984.

Le groupe de travail disciplinaire a effectué une analyse des programmes officiels qui a débouché sur une liste de 23 objectifs généraux répartis en 4 grandes catégories : connaissances, méthodologie scientifique, savoir-faire, attitudes. Le croisement entre les 23 objectifs généraux et 26 grands domaines de contenu aboutit à la définition de 598 objectifs spécifiques potentiels (tableau pp. 2 et 3). A ce niveau, il semble que l'ensemble du programme soit bien couvert. C'est à l'étape suivante que la méthodologie suivie devient plus floue. En effet, seuls certains de ces objectifs spécifiques sont opérationnalisés en exercices, les concepteurs de l'évaluation recherchant *"un équilibre entre les différents domaines de contenu et les différents objectifs de formation".*

L'évaluation de 1984 évalue 14 objectifs généraux (sur 23) décomposés en 45 objectifs spécifiques[22] (sur les 598 identifiés), à travers 30 exercices comportant de 1 à 11 items, soit 108 items au total. Remarquons que moins de 10 % des objectifs identifiés sont évalués et que, sur ceux qui le sont, 80% ne sont évalués que par un seul exercice qui ne comporte la plupart du temps qu'un petit nombre de questions.

Si l'on en reste au niveau des 4 catégories d'objectifs généraux, on note que :

- l'échantillonnage des **Connaissances** dans les domaines de contenu est le plus représentatif : 2/3 des domaines de connaissance sont évalués, avec 49 items sur les 108 de l'épreuve (soit 45%). Taux d'échantillonnage des objectifs spécifiques : 38%.
- dans la catégorie **Méthode expérimentale**, 6 objectifs généraux sur 9 sont évalués, généralement avec 1 ou 2 items. A noter qu'un objectif est privilégié, celui de *l'exploitation d'informations ou de résultats expérimentaux*, avec plus de la moitié des exercices de cette catégorie. Taux d'échantillonnage des objectifs spécifiques : 5%.
- sur les 5 objectifs de la catégorie **Savoir-faire**, deux sont évalués avec un exercice chacun, deux autres avec 4 ou 5 exercices chacun (*utilisation d'instruments de mesure, utilisation de différents langages*). Taux d'échantillonnage : 8%.
- enfin, dans la catégorie **Attitudes**, 2 objectifs généraux sur 7 sont évalués avec chacun 1 exercice. Taux d'échantillonnage : 1%.

[22] Un même exercice comportant plusieurs questions peut évaluer jusqu'à 4 objectifs.

De plus certains domaines de contenus sont évalués à travers plusieurs objectifs généraux (jusqu'à 6 pour les ions en solution à travers 8 exercices et 22 items, soit 20% de l'ensemble) alors que d'autres sont tout simplement ignorés.

On voit donc que les objectifs spécifiques sont en si grand nombre qu'ils ne peuvent matériellement être tous évalués. De surcroît, pour chaque objectif spécifique, un grand nombre d'exercices, portant sur des contenus différents de ceux retenus, pourraient être imaginés. Nous n'avons pas d'information sur les parts respectives dévolues par les concepteurs de l'épreuve aux différents objectifs généraux ou aux domaines de contenu. Seul le questionnaire aux enseignants permet d'avoir une idée de l'importance que ceux-ci attribuent à chacun des objectifs spécifiques. On remarquera cependant que les réponses ne portent que sur les objectifs définis a priori par les concepteurs. Il est donc loin d'être certain que cette épreuve soit réellement représentative (en terme de contenu mais aussi en terme de pondération de ces contenus) du programme de 3ème, encore moins de celui de l'ensemble des acquisitions antérieures. Il nous paraît donc nécessaire de définir plus précisément les critères d'échantillonnage permettant d'assurer la représentativité des items retenus.

Pour 1995 et 1999 l'évaluation repose sur les programmes de 1985 et leurs compléments. Si un tableau d'objectifs est encore fourni, leur intitulé est différent de celui utilisé en 1984, ce qui ne facilite pas les comparaisons. Alors qu'à cette époque on distinguait 4 types d'objectifs généraux, on n'en trouve plus que 2 en 1995 (connaissances et savoir-faire méthodologique et expérimental). A partir de cette date, la notion d'objectif spécifique n'est plus utilisée. Pour l'objectif général relatif aux connaissances, seules les notions à acquérir sont répertoriées sans l'utilisation d'un verbe d'action. On peut cependant reconnaître des objectifs spécifiques dans la décomposition de l'objectif général de savoir-faire selon les domaines.

2. La qualité des protocoles :

La standardisation des exercices et des consignes, tant de passation que de correction, est prise en compte par les concepteurs. On peut toutefois, s'interroger sur le respect du suivi des consignes de passation, notamment en ce qui concerne les travaux pratiques où les conditions peuvent avoir été sensiblement différentes d'un établissement à l'autre en fonction, notamment, de la disponibilité du matériel nécessaire et de la latitude donnée aux enseignants pour la constitution des groupes d'élèves et la répartition des épreuves dans le temps.

En 1999, plusieurs erreurs se sont glissées dans les cahiers des élèves, pouvant rendre incompréhensibles certains items ou pouvant troubler les élèves. La manière dont ceux-ci ont été informés (ou pas) de ces erreurs a probablement nui à la standardisation.

En ce qui concerne la fiabilité de la mesure, on a noté que la plupart des objectifs sont évalués à partir d'un seul item, qui est par nature spécifique et dont la réussite ne peut à elle seule assurer que l'objectif évalué est atteint. Il est dans ce cas problématique de généraliser la réussite (ou l'échec) à un seul item à l'acquisition de la compétence qui lui est associée.

Une faiblesse importante réside sans doute, comme nous l'avons vu dans le paragraphe précédent, dans l'échantillonnage des items, la couverture du programme étant probablement insuffisante.

Enfin, le manque d'indicateurs psychométriques (discrimination, homogénéité des items) rend difficile l'évaluation de la qualité des épreuves.

Dans la diffusion des documents au public, si en 1984 on dispose dans le «document de travail » d'une analyse détaillée des résultats de chaque exercice, on est surpris que ne soient pas publiés des résultats plus globaux concernant chaque objectif général et les quatre grandes catégories d'objectifs. En 1995, par contre, la publication de résultats détaillés n'est faite que pour quelques exemples alors que les pourcentages de réussite sont livrés pour les principaux domaines de connaissances (mais toujours pas pour les objectifs généraux). Ce manque de cohérence dans le choix des résultats publiés gêne la comparaison temporelle pour les destinataires des documents.

3. Le problème de la comparabilité des évaluations :

L'objectif de comparaison des acquis des élèves en fin de 3ème est affirmé dès 1995 et plus encore en 1999 où les documents fournis tant aux élèves qu'aux professeurs portent en titre *"Etude comparative en classe de troisième".*

Pour assurer la comparaison, il est indispensable d'avoir un nombre suffisant d'items communs aux trois évaluations. Or aucun des items de 1984 n'est repris de façon strictement identique en 1995. Certains items se ressemblent mais des modifications plus ou moins profondes les affectent en 1995. En 1999, les concepteurs laissent entendre que les 2/3 des items de l'épreuve seraient sensiblement les mêmes qu'en 1995. Une analyse détaillée montre qu'en réalité 40 % au plus des items de 1999 étaient déjà présents dans l'épreuve de 1995. Réciproquement, compte tenu du nombre plus élevé d'items en 1995, cette épreuve ne comporte que 28% des items de 1999. En outre, la comparaison est rendue difficile par la longueur variable de l'épreuve, très différente entre les 3 années, puisqu'elle passe de 108 items en 1984 à 65 en 1999.

Une analyse a posteriori révèle d'autre part une proportion plus importante d'items difficiles (échoués par plus de 10 % des élèves) en 1999 qu'en 1984. Inversement, la proportion d'items faciles est plus importante en 1984 qu'en 1999. Ceci peut contribuer à une baisse de la réussite globale cette année-là.

La standardisation des épreuves doit également être identique chaque année, et en particulier les conditions de passation et de cotation. Concernant les conditions de passation, les exercices se voyaient attribuer un temps précis en 1984 alors qu'à partir de 1995, un temps global est alloué à l'ensemble de l'épreuve. En ce qui concerne la cotation, on note également quelques différences. Il en est ainsi, entre autres, d'un item relatif à la propagation de la lumière, où un code supplémentaire est prévu en 1999, rendant la comparaison de la réussite difficile, une réponse admise en 1995 ne l'étant plus en 1999 (baisse de réussite de 40%). Des erreurs subsistent également dans le guide de correction en 1999, certains renvois étant incompréhensibles, car identifiant des items selon leur numérotation de 1995 qui n'était plus la même en 1999. Ces erreurs ont également pu biaiser la correction.

Compte tenu de ces remarques, il nous paraît donc illégitime de comparer les taux de réussite des évaluations des 3 années. De même, la comparaison item par item est rendue délicate puisque ceux-ci ne sont pas placés de façon identique dans les trois évaluations. Par exemple, certains peuvent une année être précédés par d'autres du même type, avec un effet d'entraînement possible, alors qu'une autre année, ce ne sera pas le cas. Enfin, nous avons vu que la comparaison de la réussite aux objectifs généraux est rendue difficile par l'absence d'identité repérable de ces objectifs d'une année à l'autre.

D'une façon plus générale, on peut se demander quel sens véritable il y a à comparer les résultats des deux dernières évaluations avec ceux de 1984 dans la mesure où le temps alloué à l'étude des Sciences physiques a été divisé par deux entre 1984 et 1995. En effet les élèves évalués à partir de 1995 n'ont pas eu d'enseignement dans cette discipline en 6ème et 5ème.

ANNEXE 9 : LIMITES DES COMPARAISONS TEMPORELLES A PARTIR DES ITEMS COMMUNS : L'EXEMPLE DES PROTOCOLES DE MATHEMATIQUES

Nous présentons quelques résultats de nos analyses sur la comparaison des performances aux protocoles de mathématiques en classe de troisième pour les évaluations bilan de 1984, 1990 et 1995 (Bonora et Vrignaud, 1997). Nous avons intégré aux données présentées les éléments du protocole 1999.

Nous avons cherché à mettre en œuvre les méthodes psychométriques adaptées pour l'étude de l'évolution temporelle sur les protocoles de mathématiques des quatre dispositifs successifs. Nous avons présenté dans le corps du texte les biais communs à l'ensemble des disciplines (biais d'échantillonnage, d'administration), nous nous attacherons, ici, à l'étude des items communs dans les protocoles de mathématiques. Comme nous l'avons signalé dans le corps du texte, la comparaison de la réussite à un item isolé est dénuée de sens. Il est nécessaire, pour cette comparaison, de construire à partir des items communs une échelle de compétence dans une métrique commune. On peut utiliser différents modèles statistiques (voir Kolen et Brennan, 1995, pour une présentation générale ; Bonora et Vrignaud, 1997, pour un exemple de mise en œuvre sur les protocoles de mathématiques de la DEP) pour effectuer cette opération – dite paralléllisation, en anglais *equating* –. Cette procédure requiert que les items communs soient fiables, représentatifs du domaine, placés à des positions similaires dans les protocoles.

Nombre d'items communs entre les 4 protocoles

Dans une perspective comparative, la DPD avait inclus des exercices communs entre les protocoles successifs. Ces exercices étaient retenus en fonction de différents critères, en particulier que leur contenu et leur présentation étaient restés accessibles à des élèves cinq ans plus tard. Certains de ces exercices ont malheureusement été parfois légèrement modifiés, faisant alors perdre *ipso facto* aux items leur fonction d'étalon de mesure. Nous avons pu ainsi débusquer une dizaine de "pseudo items communs", que nous avons dû éliminer de la liste.

Par exemple, un exemple typique d'altération de la démarche requise pour aboutir à une bonne réponse est représenté par l'exercice suivant. Il s'agit dans tous les cas de calculer la hauteur d'un immeuble, connaissant la distance de l'observateur, sa taille, et l'angle sous lequel il voit cet immeuble. La procédure de résolution observable est constante à travers les deux versions de l'item. Cependant l'altération de la démarche psychologique requise de l'élève entre les versions 1984 et 1990 a une double origine :

- En 1984 le problème est posé en termes abstraits et de façon indirecte ("la figure schématise la détermination de la hauteur d'un immeuble..."), alors qu'en 1990 et 1995, il l'est de façon plus "concrète", et après une mise en situation ("un observateur est situé à 80 m d'un immeuble. Ses yeux se trouvent à 1,63 m du sol ...").
- En 1990 (et 1995) on fournit la valeur de l'angle de visée, alors qu'en 1984 on fournit la valeur de la tangente (ainsi que les valeurs des sinus et cosinus).

Une fois éliminés ces « pseudo-items communs », on dispose de 62 items qui sont communs à deux, aux trois ou aux quatre dispositifs, ce qui, de prime abord, paraît être un effectif suffisant pour permettre d'envisager des procédures d'*equating* entre ces cohortes. Tout d'abord, récapitulons-les au plan des cohortes. On a la répartition suivante :

Effectifs des items communs aux quatre protocoles

	1984	**1990**	**1995**
1990	38 (22)		
1995	6 (6)	21 (7)	
1999	5 (5)	20 (16)	29 (14)

Les chiffres entre parenthèses indiquent le nombre d'items qui devront être éliminés pour les opérations de parallélisation.

Notons tout d'abord que le total calculable ici (n = 119) est supérieur à l'effectif global rapporté précédemment (n = 62), simplement parce que les items communs à plus de deux cohortes figurent, de ce fait, parmi les effectifs observés dans chacune des cases correspondantes.

Les situations qui paraissent les plus favorables en raison de ces effectifs d'items communs sont celles de l'*equating* 1984-90 et 1995-1999, car les effectifs y sont nettement les plus élevés. Par ailleurs, un *equating* effectué directement entre les cohortes "extrêmes" (1984 et 1999 ; 1984-1995) reposerait sur des bases fragiles (respectivement 5 et 6 items seulement).

Pour la fiabilité de l'opération de parallélisation, il est nécessaire de retirer les items présentant un fonctionnement défectueux, soit parce qu'ils sont inconsistants (discrimination insuffisante), soit qu'ils présentent un fonctionnement différentiel (influencés par les changements de contextes). Le nombre d'items éliminés est placé entre parenthèses dans chaque case. On voit que le nombre d'items communs utilisables se réduit de manière drastique. Signalons qu'aucun item n'apparaît fiable pour procéder à des comparaisons directes sur les cohortes extrêmes (1984 et 1995 ou 1999).

Représentativité des items communs

Dans l'optique où on cherche à évaluer les changements dans les compétences des élèves au niveau d'une discipline, il est indispensable que les items communs soient représentatifs du contenu de la discipline.

Nous avons utilisé, pour établir cette répartition (tableau 2), la nomenclature mise en place pour les dispositifs de 1990, 1995 et 1999, qui n'a pas varié entre les trois moments. Afin de permettre la comparaison, entre la répartition des items communs et de l'ensemble, nous avons effectué un reclassement des items de l'ensemble des cahiers de 1984, en se basant sur cette même nomenclature. Si pour chacun des protocoles, chaque objectif est représenté par au moins une dizaine d'items, il n'en est plus de même au niveau des items communs. On remarquera la nette sous-représentation de « gestion des données » entre 1990 et 1995. Il apparaît des différences de proportion des différents domaines entre les protocoles et la partie commune. Ces différences de représentativité sont encore accentuées lorsqu'on examine les items utilisables. Une fois retirés les items méthodologiquement inutilisables, certains objectifs ne comportent aucun item commun. Nous avons indiqué le nombre d'items qui devront être éliminés de la suite des procédures pour la comparaison 1995/1999. On voit que près de la moitié des items ne peuvent donner d'indications fiables et que les performances aux objectifs « gestion de données » et « géométrie » ne peuvent plus être valablement comparés.

Nombre d'items par protocole et nombre d'items communs à deux évaluations successives répartis par objectifs.

La taxonomie des objectifs commune à 1990, 1995 et 1999 a été utilisée, les exercices de 1984 ont été reclassés selon cette typologie (Bonora et Vrignaud, 1997).

	1984	**comm 84/90**	**1990**	**comm. 90/95**	**1995**	**Comm. 95/99**	**1999**
travaux numériques	39	12	31	10	19	14 (6)	15
gestion de données	13	12	18	1	8	5 (2)	7
géom. (plan et espace)	45	14	44	10	28	10 (6)	10
TOTAL	97	38	93	21	55	29 (14)	32

Les chiffres entre parenthèses, dans la colonne des items communs 95/99, indiquent le nombre d'items qui devront être éliminés pour les opérations de parallélisation.

Position des items communs

L'équité de la comparaison des réussites entre items communs pour deux cohortes différentes suppose notamment que chaque item considéré soit situé à peu près au même rang dans l'ordre de passation par les deux cohortes, afin que les sujets des deux groupes soient dans des conditions similaires (condition d'habituation à la situation, et/ou de fatigue pour les items en fin d'épreuve). On constate qu'en règle générale le positionnement des items communs est assez différent d'un dispositif à l'autre.

Synthèse

Le nombre faible d'items communs utilisables pour les procédures statistiques, leur inégale répartition – voire leur absence – selon les objectifs évalués, leur positionnement souvent différent à l'intérieur des protocoles ne permettent pas de mettre en œuvre de manière fiable des procédures de comparaisons entre les quatre protocoles. Nous avons observé des problèmes similaires pour les autres disciplines étudiées (allemand, histoire-géographie, physique).