Évaluation des acquis des élèves à l'école obligatoire

L'évaluation cantonale et commune : état de situation du dispositif existant, points forts et points faibles, perspectives d'avenir

Anne Soussi, Edith Guilley, Ninon Guignard et Christian Nidegger

Mai 2009

REPUBLIQUE ET CANTON DE GENEVE
POST TENEBRAS LUX

Service de la recherche en éducation

Évaluation des acquis des élèves à l'école obligatoire

L'évaluation cantonale et commune : état de situation du dispositif existant, points forts et points faibles, perspectives d'avenir

---

Anne Soussi, Edith Guilley, Ninon Guignard et Christian Nidegger

Mai 2009

Compléments d'information :
Anne Soussi
Tél. +41/0 22 546 71 39
anne.soussi@etat.ge.ch

Edith Guilley
Tél. +41/0 22 546 71 51
edith.guilley@etat.ge.ch

Ninon Guignard
Tél. +41/0 22 546 71 25
ninon.guignard@etat.ge.ch

Christian Nidegger
Tél. +41/0 22 546 71 19
christian.nidegger@etat.ge.ch

Responsable de l'édition :
Narain Jagasia
Tél. +41/0 22 546 71 14
narain.jagasia@etat.ge.ch

Internet :
http://www.ge.ch/sred

Diffusion :
Service de la recherche en éducation (SRED)
12, quai du Rhône - 1205 Genève
Tél. +41/0 22 546 71 00
Fax +41/0 22 546 71 02

Document 09.013

Le contenu de ce document n'engage que la responsabilité du Service de la recherche en éducation

Remerciements

Nos remerciements s'adressent à toutes les personnes qui ont bien voulu nous consacrer du temps, répondre à nos questions ou nous fournir les informations nécessaires :

– le secrétaire général, M. Frédéric Wittwer ;

– le secrétaire adjoint, M. Renato Bortolotti ;

– les deux directrices de l'enseignement à la direction de l'enseignement primaire et du cycle d'orientation, Mmes Thérèse Guerrier et Bernadette Badoud-Volta, ainsi que Mme Isabelle Nicolazzi qui a remplacé cette dernière ;

– le coordinateur des épreuves cantonales, adjoint à la direction de l'enseignement primaire, M. Ladislas Ntamakiliro ;

– le responsable du secteur de l'évaluation commune au cycle d'orientation (à la retraite depuis juillet 2008), M. François Bugniet, ainsi que la personne qui lui a succédé, M. Philippe Rouget ;

– le responsable de la préorientation au cycle d'orientation, M. Emiel Reith ;

– les concepteurs des épreuves du CEFEP :

  - en français : Mmes Muriel Wacker, Françoise Vodoz et M. Denis Métroz ;

  - en allemand : Mme Lotti Kuster ;

  - en mathématiques : Mmes Blandine Choquet et Muriel Corthésy, MM. Eric Burdet et Jean-Pierre Bugnion ;

  - ainsi que Mme Christiane Jeannet du SEDEV ;

– les commissaires des épreuves du CO :

  - en français : Mmes Marianne Moser (PG), Emmanuelle Tarazzi et Caroline Salvi, MM. Dominique Pellizari et Bernard Pinget ;

  - en allemand : Mmes Christiane Winter (PG), Linda Souchard, Carole de Jong, Heike Vuillemin Beucher, Svetlana Stevanovic-Zaric et Gabriele Zimmermann ;

  - en mathématiques : Mmes Rita Joye-Bortolotti et Josiane Bloechlinger, MM. Claude Lecoultre (PG), Philippe Dubath (PG), Sébastien Archimède, Pierrick Dudognon et Andrea Reuben ;

  - en physique : MM. Christian Colongo et Jacques Bochet (PG) ;

– les enseignants des écoles primaires et des établissements du cycle d'orientation ainsi que leurs directions.

Nous remercions également vivement M. Daniel Bain pour sa collaboration, sa disponibilité et ses précieux conseils dans l'analyse docimologique, ainsi que Mme Lucie Mottier-Lopez de la FPSE pour ses conseils et les références bibliographiques qu'elle nous a fournies.

Merci également à notre collègue, Pierre-Alain Wassmer, pour son aide dans la récolte d'information concernant l'estimation des coûts des évaluations cantonales et communes.

Nos remerciements vont aussi à toutes les personnes qui ont retranscrit les entretiens : Mmes Monique Butzer, Dominique Chenu, Maria El-Hindi, Elda Anting et M. Bernard Engel, sans lesquels il n'aurait pas été possible de réaliser ce travail.

Table des matières

Résumé

Dans la perspective de la réorganisation de l'enseignement obligatoire en une seule direction pour le primaire et le cycle d'orientation (CO), et des projets d'évaluation externe nationaux et romands, un mandat a été confié au SRED pour réaliser un état de situation du dispositif d'évaluation des acquis des élèves existant (épreuves cantonales et communes) en mettant en évidence ses points forts et ses points faibles et donner des éléments de réflexion en vue de la nouvelle organisation en tenant compte des futures évaluations prévues au niveau national et romand.

La présente étude a été réalisée en deux phases :

- la première phase a été consacrée à une analyse des évaluations externes en Europe et en Suisse romande, à des entretiens avec les principaux acteurs impliqués dans l'élaboration des épreuves cantonales et communes (autorités, directions de l'enseignement, responsables de l'évaluation dans les deux niveaux d'enseignement ainsi que les concepteurs de ces épreuves), à des analyses docimologiques de la qualité technique des épreuves ;
- la seconde phase a permis de compléter l'analyse des épreuves et de récolter l'avis d'un petit échantillon d'enseignants de l'école primaire et du CO, en tant que principaux utilisateurs de ces épreuves. Une analyse de parcours d'élèves de la 7e à la 9e en fonction de leurs résultats aux épreuves cantonales de 6P a également été réalisée.

Les points forts et les points faibles des évaluations cantonales et communes

Les dispositifs d'évaluation existants aux deux niveaux d'enseignement sont organisés de manière assez différente : à l'école primaire les épreuves sont élaborées par des formateurs alors qu'au CO elles sont conçues par des commissaires d'épreuves qui sont des enseignants dégrevés.

L'analyse des points forts et faibles en fonction de trois critères (qualité technique, utilité et efficience/efficacité) a permis de relever de nombreux éléments :

Au niveau de la qualité technique

Dans les deux ordres d'enseignement, la plupart des acteurs interrogés relèvent une bonne qualité des épreuves, confirmée par l'analyse docimologique de quelques épreuves en mathématiques et en français qui conclut à une bonne fiabilité pour la certification des élèves et la régulation de l'enseignement. A l'école primaire, les *points forts* relevés par l'ensemble des acteurs concernent différents éléments tels que l'existence de prétests, de tables de spécification permettant d'identifier précisément les objectifs évalués, la qualité des explications fournies au niveau des conditions de passation et de correction, l'expérience des concepteurs d'épreuves. Au CO, les *points forts* sont souvent liés aux disciplines : stabilité du contenu et du niveau de difficulté (p. ex. en mathématiques), bonnes conditions de passation et de correction, clarté de l'épreuve et des consignes de correction (p. ex. en anglais et en physique).

Certains *points faibles* mentionnés concernent les épreuves du primaire comme celles du CO : différences d'approches entre les disciplines (et nécessité d'harmoniser ces approches), fiabilité insuffisante pour évaluer des connaissances et compétences détaillées. L'analyse plus détaillée des résultats, voire le recours systématique aux prétests, est souhaitée par les différents acteurs (au primaire, les prétests sont systématiques mais servent seulement à tester la formulation des consignes et à déterminer la table de spécification). Les autorités relèvent un certain manque de standardisation

des épreuves et de cohérence tout au long de la scolarité obligatoire et estiment nécessaire de constituer des équipes pluridisciplinaires, comportant des personnes avec de bonnes connaissances du programme et d'autres avec des compétences liées à la construction d'épreuves et leur analyse. A l'école primaire, l'ensemble des acteurs interrogés relèvent l'existence de biais liés aux conditions de passation et de correction (les enseignants en charge des élèves assurant eux-mêmes ces deux opérations), la couverture incomplète du programme par les épreuves, l'absence d'enseignants dans les commissions d'épreuves. Le manque de stabilité d'une année à l'autre au niveau du contenu des épreuves et parfois le niveau d'exigences estimé peu adapté (p. ex. pour certaines épreuves de 2P) sont mis en avant par les enseignants.

Au CO, on déplore le manque de prétest pour certaines disciplines, faute de temps. Les différents acteurs mentionnent également : l'absence de tables de spécification dans les épreuves de certaines disciplines, de probables biais de correction, le manque de stabilité et la formation des concepteurs jugée parfois insuffisante (notamment lorsqu'on introduit de nouveaux objets), la manière parfois inadéquate de prendre en compte le programme et de tester les objectifs (en particulier pour évaluer les compétences complexes comme la production écrite), la façon de constituer les barèmes et de fixer le seuil de réussite dans la plupart des disciplines. Une remarque générale mentionnée par l'ensemble des personnes interrogées au CO se rapporte à la difficulté à prendre en compte des élèves de niveaux de compétence différents au travers d'une même épreuve (avec des barèmes différenciés) étant donné le postulat des mêmes objectifs pour tous.

Au niveau de l'utilité

Un des principaux *points forts* de ces évaluations relevé par l'ensemble des acteurs interrogés aux deux niveaux d'enseignement est d'être communes à tous les élèves d'une volée et d'avoir des exigences communes pour une discipline donnée. Ces épreuves permettent de réguler l'enseignement et d'unifier les pratiques.

Les objectifs de ces épreuves définis institutionnellement sont confirmés par l'analyse docimologique de quelques épreuves en français et en mathématiques : la certification des élèves et la régulation de l'enseignement.

Parmi les *points faibles* relevés, il ressort que parfois ces épreuves seraient utilisées pour répondre à trop de besoins : outre la vérification de l'atteinte des objectifs du programme par les élèves, elles joueraient également un rôle dans le pilotage du système. D'autres acteurs regrettent l'absence d'une évaluation-système.

L'analyse et l'exploitation des résultats sont estimées insuffisantes par la majorité des acteurs.

Des effets pervers des épreuves sont également mentionnés, notamment au CO tels que la tendance au bachotage et la réduction du programme et de l'enseignement au contenu des épreuves.

Au primaire en particulier, les enseignants déplorent l'importance de ces épreuves pour les élèves et leurs parents, la lourdeur des épreuves et le stress engendré notamment pour les élèves de 2P. Au CO, dans certaines disciplines, on estime que le type d'évaluation proposé est trop différent de ce qui se fait habituellement en classe, notamment dans le regroupement B.

Au niveau de l'efficacité/efficience

De manière générale, le dispositif d'évaluation cantonale/commune peut être considéré comme efficace car il répond globalement aux objectifs définis. Toutefois, au niveau de l'efficience, les informations récoltées (notamment les avis des acteurs) ne permettent pas de tirer des conclusions par rapport à ce critère.

L'analyse des parcours d'élèves de la 7^{e} à la 9^{e} en fonction de leurs résultats aux épreuves cantonales de 6P montre que ces épreuves ont une bonne prédictivité même si ce n'est pas leur objectif prioritaire. Une part très importante des élèves qui réussissent les épreuves en 6P se retrouvent encore en 9^{e} dans

le regroupement A. L'analyse docimologique met également en évidence une bonne cohérence entre la réussite aux épreuves de 6P et de 7e (cas des mathématiques).

Un *point positif* relevé par la majorité est l'existence au CO d'un outil informatique précieux, EVACOM qui permet la saisie des résultats aux épreuves.

Parmi *les points faibles* relevés, on trouvera la place jugée trop importante de ces épreuves dans les carnets.

La surcharge de travail liée à la correction et la saisie des résultats a parfois été relevée au CO.

Réflexions en vue d'une nouvelle organisation de l'évaluation cantonale et commune

Les différentes analyses ont permis de dégager des éléments de réflexion en vue d'une nouvelle organisation de l'évaluation cantonale ou commune, qui s'organisent autour des points suivants :

Objectifs et attentes des évaluations externes

Il y a un consensus, en tout cas formel, sur les objectifs des épreuves cantonales et communes dans les deux niveaux d'enseignement : elles servent à vérifier l'atteinte des objectifs et participent à la certification mais le rôle est un peu différent selon les moments (p. ex. statut de la 2P ; 6P et 9e). Au CO les épreuves participent également à l'orientation et la sélection.

Les différentes évaluations devraient jouer des rôles différents : les tests HarmoS et l'enquête PISA visent plutôt l'évaluation du système et le monitorage, les épreuves cantonales ou communes ont plutôt pour but de certifier les élèves et de vérifier l'atteinte des objectifs. Quant aux objectifs des épreuves de référence communes romandes, ils ne sont pas pour l'instant bien définis (certification, diagnostique ou bilan).

Contenu des épreuves et manière d'évaluer

Il est nécessaire de réfléchir sur différents éléments tels que le format des questions et le choix des objectifs, la représentativité des objectifs du plan d'études (prise en compte des objectifs du Plan d'études et leur pondération dans l'épreuve), la stabilité des épreuves (mêmes domaines pris en compte d'une année à l'autre dans un souci de comparabilité), la manière de mesurer des compétences complexes (notamment en langues, la production écrite en particulier).

Procédures pour atteindre ces objectifs

Il est nécessaire de généraliser les prétests dans toutes les disciplines, non seulement pour vérifier la compréhension des consignes mais également pour fixer des seuils de réussite et contrôler la qualité technique des épreuves (p. ex. validité, fiabilité).

La prise en compte des différents niveaux de compétences des élèves (regroupements différenciés) reste un point crucial et délicat où la réflexion est encore à poursuivre.

Administration, utilisation et exploitation de l'évaluation externe

Le fait de croiser les classes (et leurs enseignants) permettrait de diminuer les biais liés aux conditions de passation et de correction.

L'analyse des résultats grâce à leur saisie sur EVACOM pourrait être détaillée par item en fonction des besoins en vue de la régulation de l'enseignement et des programmes.

Compétences et ressources nécessaires

Différentes compétences sont nécessaires et les équipes pourraient être pluridisciplinaires comportant des personnes avec une bonne connaissance du programme et des experts en docimologie. Un autre groupe pourrait être mis sur pied pour vérifier l'adéquation de la sélection des objectifs du Plan d'études et leur représentativité dans l'épreuve cantonale. Faut-il que l'organisation des épreuves cantonales et communes soit transversale ? On pourrait imaginer que des personnes différentes élaborent les épreuves à l'école primaire et au CO étant donné que les approches définies dans les plans d'études et les enseignants ne sont pas les mêmes aux deux niveaux d'enseignement mais qu'ils collaborent (surtout dans les disciplines communes aux deux niveaux d'enseignement) en vue d'une harmonisation ou d'une certaine cohérence serait utile. Les experts en docimologie pourraient être, quant à eux, communs aux deux niveaux.

Les ressources nécessaires seront probablement en augmentation dans un premier temps avec la généralisation des prétests, le développement des analyses mais à plus long terme on peut supposer un gain dû à la professionnalisation des concepteurs. Par ailleurs, l'analyse des liens avec les évaluations des enseignants est également à développer.

Introduction

L'évaluation est un phénomène complexe de manière générale et dans le domaine de l'éducation en particulier. Quand on parle d'évaluation, on suppose un jugement orienté vers un but et qui va donner lieu à une prise de décision. En 1986 déjà, Cardinet la définissait comme « une démarche d'observation et d'interprétation des effets de l'enseignement, visant à guider les décisions nécessaires au bon fonctionnement de l'école » (p. 13). L'évaluation peut se situer à différents niveaux : élève, classe, enseignant, établissement ou encore système de formation. Elle peut également être de type interne, c'est-à-dire réalisée par l'enseignant pour ses élèves, ou de type externe, c'est-à-dire réalisée par d'autres personnes que l'enseignant lui-même. Les évaluations externes peuvent prendre différentes formes : épreuves de référence ou cantonales réalisées par des commissions comprenant le plus souvent des enseignants, des formateurs et d'autres membres de l'institution scolaire. Elles peuvent également relever d'organismes ou d'instances externes (p. ex. les enquêtes internationales à large échelle comme PISA). Les futurs tests HarmoS en tant que projet national, conçus probablement par des organismes externes composés d'experts, peuvent également entrer dans cette catégorie.

Pour Weiss (2002), on peut distinguer d'une part les niveaux, interne ou externe mais également celui des logiques, c'est-à-dire l'objectif visé, logique de l'apprenant ou celle de l'organisation. Le schéma suivant résume ces différentes situations. Selon lui, « l'évaluation en milieu scolaire remplit de multiples fonctions, régulation des apprentissages, information des familles, pronostic de la réussite, bilan du système de formation » (2002, p. 2).

Tableau 1. Les différentes logiques de l'évaluation

	Logique de l'apprenant	Logique de l'organisation
Évaluations externes	- épreuves diagnostiques pour l'évaluation des apprentissages des élèves - épreuves bilan pour la certification des apprentissages des élèves	- évaluation bilan des systèmes de formation (cantons, pays) - évaluation pour la certification des organisations apprenantes (école, établissement)
Évaluations internes	- évaluation formative des apprentissages des élèves - évaluation bilan-sommative des apprentissages des élèves	- évaluation régulatrice autogérée d'une organisation apprenante

Ces différentes fonctions sont parfois en concurrence, l'un des objectifs étant officiel, les autres officieux. Dans la majorité des systèmes scolaires, on trouve les deux types d'évaluations internes et externes. Dans le mandat reçu par le SRED qui porte sur l'évaluation des acquis à l'école obligatoire, la centration se fera sur les évaluations « externes » de type cantonale ou commune. Ces évaluations ne sont pas tout à fait externes selon certaines classifications (notamment Hanay et Madaus, 1986, cités par Monseur et Demeuse, 2005) car si elles sont élaborées par des personnes externes (autres que les enseignants des élèves évalués), elles sont administrées et corrigées par les enseignants (titulaires des classes concernées). Nos évaluations cantonales et communes pourraient être qualifiées de mixtes. Elles se situent plutôt du côté de la logique de l'apprenant même si, comme on le verra plus loin, elles peuvent parfois servir à l'organisation également.

Le mandat et l'origine du projet

Le SRED a reçu un mandat du secrétariat général visant à réaliser, avec l'aide des directions générales de l'école primaire (DGEP) et du cycle d'orientation (DGCO), une évaluation du dispositif d'évaluation des acquis des élèves dans l'enseignement obligatoire. Plus précisément, il est demandé au SRED :

- *d'établir un inventaire des dispositifs d'évaluation existants qui s'adressent aux élèves, aux établissements scolaires et au système scolaire genevois, en mettant en évidence l'articulation entre les différents dispositifs. Cet inventaire doit en particulier aborder les points suivants :*

  a) les objectifs et les contenus des différentes évaluations ;

  b) les modalités de conception, de réalisation, de traitement et d'analyse des évaluations ;

  c) l'utilisation de ces évaluations.

- *d'analyser les points forts et les opportunités d'amélioration des dispositifs d'évaluation existants ;*

- *de mettre cela en perspective avec les nouveaux dispositifs d'évaluation prévus sur les plans suisse (HarmoS : tests de référence des standards) et régional (épreuves de référence du plan d'études romand) ;*

- *de proposer une organisation transversale d'évaluation des acquis des élèves dans l'enseignement obligatoire qui réponde aux critères de qualité des démarches évaluatives, qui soit en cohérence avec les finalités des évaluations des élèves prévues sur les plans régional et national et qui soit compatible avec la mise en place d'une direction de l'enseignement obligatoire ;*

- *d'établir un plan de mise en œuvre.*

(cf. note du 9.4.08, de M. F. Wittwer, secrétaire général du DIP).

L'origine du mandat peut être attribuée principalement à deux éléments :

- d'une part, l'*Accord intercantonal sur l'harmonisation de la scolarité obligatoire* HarmoS qui définit les finalités essentielles de la scolarité obligatoire et prévoit l'évaluation régulière des performances de l'école obligatoire au moyen de tests de référence qui vérifient l'atteinte des standards nationaux de formation et la Convention scolaire qui prévoit notamment la mise en place d'épreuves romandes communes pour vérifier l'atteinte des objectifs du plan d'études romand (PER) ;
- d'autre part, la mise en place d'une direction générale de l'enseignement obligatoire qui est prévue à la rentrée 2009. Cette nouvelle organisation rend nécessaire la recherche d'une certaine cohérence au niveau des dispositifs d'évaluation existants dans le système d'enseignement obligatoire.

Le choix a été fait par les différentes personnes impliquées dans l'élaboration du mandat (secrétariat général, DGEP et DGCO, direction du SRED) de mettre l'accent sur l'évaluation externe, et plus particulièrement les épreuves cantonales et communes. Il a été décidé de profiter de l'occasion de mise à plat des dispositifs existants pour s'intéresser également de manière succincte à la préorientation, à laquelle tous les élèves de 6P participent et qui est une autre forme d'évaluation externe.

Le mandat s'est déroulé en deux phases. Au départ, il était prévu sur une période très courte : initié à la fin mai, le rapport devait être rendu à la fin septembre. Étant donné ces délais très courts impartis, nous avons choisi, dans un premier temps, plutôt d'approfondir certains points : les opinions concernant le dispositif existant des membres du secrétariat général, des deux directrices de l'enseignement et des concepteurs d'épreuves dans un certain nombre de disciplines à l'école primaire et au CO ainsi que l'analyse de la validité et de la fiabilité de quelques épreuves de français et de

mathématiques. Dans un second temps (de novembre à mi-février), afin de cerner également l'opinion des principaux utilisateurs que sont les enseignants, un échantillon d'enseignant-e-s de l'école primaire et du CO ont été interrogés à propos des épreuves cantonales ou communes. Les analyses (docimologiques et qualitatives) de ces épreuves se sont poursuivies ainsi qu'une observation des liens entre les résultats aux épreuves en 6P et l'orientation des élèves de la 7e à la 9e année. L'estimation des coûts des épreuves a fait l'objet d'une première tentative avec la collaboration des deux directions de l'enseignement concernées. Les informations récoltées se sont avérées parcellaires et peu exploitables. Ce volet devra être développé dans le cadre d'un mandat complémentaire.

Description de la méthode

La section I sera consacrée à un bref survol de la situation européenne et romande en matière d'évaluation externe (analyse documentaire).

Dans la section II, la situation à Genève sera appréhendée de plusieurs manières :

- un recueil de documents officiels portant sur les évaluations cantonales ou communes ;
- des entretiens avec différents acteurs concernés : secrétariat général, directrices de l'enseignement des deux niveaux d'enseignement de l'école obligatoire, coordinateur ou responsable de l'évaluation cantonale ou commune, commissions de concepteurs de ces épreuves (formateurs du CEFEP en français, allemand, mathématiques et évaluation pour l'école primaire et commissions d'enseignants chargées de l'élaboration des épreuves en français, mathématiques, allemand et physique pour le CO[1]), ce qui a représenté en tout 15 entretiens de 1h30 à 2h. Ils ont pour but de décrire le plus précisément possible la situation existante dans un premier temps, puis de mettre en évidence les points forts et les points faibles du dispositif actuel de manière générale et en fonction de critères spécifiques. Pour identifier les points forts et les points faibles du dispositif existant, nous avons construit une grille et considéré trois grandes catégories de critères repris et adaptés de grilles existantes (Stufflebeam 1974, 2003 ; Shepard, 1977 ; Scriven, 1974 et Stake, 1969). Il s'agit des critères d'*adéquation* et de *qualité technique* (validité interne, externe, fidélité/fiabilité, etc.), d'*utilité* (pertinence par rapport aux publics visés, crédibilité auprès des publics visés, diffusion des résultats, etc.) et d'*efficience/efficacité* (temps de travail, postes, budget, etc.). Deux thématiques sont également abordées dans les entretiens : l'opinion des autorités, directions d'enseignement et concepteurs des épreuves concernant une organisation idéale de l'évaluation cantonale/commune et la place des différentes évaluations au plan national (tests HarmoS), romand (épreuves de référence) et cantonal (évaluations cantonales ou communes). Dans une seconde phase, des entretiens ont également été réalisés avec un échantillon d'enseignants sélectionnés par école à l'école primaire et par discipline et établissement au CO concernant leur opinion à propos des évaluations cantonales ou communes ;
- des analyses statistiques permettant d'évaluer la qualité technique de certaines de ces épreuves (déjà administrées) : fiabilité/fidélité, validité, etc. Ces analyses permettent de compléter le point de vue des acteurs concernant le premier type de critères, à savoir la qualité technique ;
- une analyse qualitative visant à comparer des épreuves de 6P et de 7e dans deux domaines, le français et les mathématiques, permet de donner un autre éclairage ;

[1] Au CO, de nombreuses disciplines (sept) font l'objet d'épreuves communes à la fin de chaque degré ou pour l'un ou l'autre des trois degrés. Il n'a pas été possible d'effectuer des entretiens avec l'ensemble des commissaires chargés des épreuves communes dans toutes les disciplines. Nous en avons sélectionné quelques-unes, notamment celles communes aux deux niveaux d'enseignement (français, mathématiques et allemand) et en avons ajouté une, la physique, qui est assez exemplaire en matière d'épreuve commune.

- enfin, une observation des parcours d'élèves ayant passé les épreuves de 6P (français I, français II et mathématiques) a été effectuée pour voir le lien entre les résultats à ces épreuves et l'orientation des élèves de la 7e à la 9e et déterminer la prédictivité de ces épreuves.

Une synthèse finale permettra de confronter le point de vue des différents acteurs concernant les points forts et les points faibles avec l'analyse de la qualité technique des instruments analysés. Cette synthèse devrait contribuer à proposer des pistes en vue d'une nouvelle organisation de l'évaluation commune/cantonale au sein de la scolarité obligatoire.

Présentation du rapport

Une première section traite de la situation en Europe et en Suisse romande.

Une seconde section, plus conséquente, aborde différents chapitres. Dans le **premier chapitre** on rappelle les objectifs officiels des évaluations cantonales et communes. Le **deuxième chapitre** est consacré au point de vue des autorités, des directions d'enseignement et des concepteurs des évaluations cantonales et communes (points forts et points faibles, notamment) mais concerne également la structure du dispositif actuel et l'organisation idéale. Dans le **troisième chapitre**, c'est le point de vue des enseignants qui est abordé. Le **quatrième chapitre** rend compte d'analyses docimologiques d'épreuves de l'école primaire et du CO du point de vue de leur validité, de leur fiabilité et de leur stabilité. Une comparaison d'épreuves de 6P et de 7e en français et en mathématiques est également effectuée pour montrer les différences et les ressemblances des conceptions des deux niveaux d'enseignement. Le **cinquième chapitre** illustre par des parcours d'élèves de la 7e à la 9e les liens entre la réussite aux épreuves cantonales de 6P et l'orientation des élèves au CO. Les **chapitres 6** et **7** abordent d'autres évaluations externes : la préorientation, les tests HarmoS, les épreuves de référence romandes ainsi que l'enquête PISA en essayant de situer leur place et celles des évaluations cantonales ou communes genevoises.

Enfin, une synthèse résume les points les plus saillants et apporte une réflexion sur l'organisation du dispositif d'évaluation cantonale/commune dans la future organisation de l'enseignement obligatoire.

I. La situation en Europe et en Suisse romande

1. La situation en Europe

Compte tenu des informations trouvées notamment dans la base de données européenne Eurybase, nous allons, dans la mesure du possible, décrire la situation européenne concernant l'évaluation externe. L'évaluation externe a pris un essor important en Europe et en Amérique du Nord depuis une bonne dizaine d'années. Elle est souvent associée aux besoins d'*accountability* (rendre compte) venant d'Amérique du Nord et à l'autonomie des écoles. D'autres courants sont également en vogue dans bon nombre de pays européens et en Amérique du Nord concernant les compétences (socles de compétences) et les standards de formation ou de performance.

Concernant l'évaluation externe, on peut globalement répartir les différents pays d'Europe en trois groupes :

- ceux qui ont, à côté de l'évaluation interne pratiquée par les enseignants, une évaluation externe (sous forme d'épreuves à différents moments du cursus) qui leur permet d'évaluer les acquis des élèves ou de piloter le système ;
- ceux qui n'ont pas à proprement parler d'épreuves externes mais qui ont mis au point un dispositif de critères pour aider les enseignants à évaluer l'atteinte des objectifs (p. ex. atteintes déclinées en niveaux dans le curriculum) ou à proprement parler des standards de performance que les élèves sont censés atteindre ;
- ceux qui utilisent uniquement l'évaluation interne de l'enseignant pour certifier les élèves et qui n'ont pas à proprement parler de dispositif « standardisé » pour attester de l'atteinte des objectifs des curricula.

L'évaluation externe peut se présenter de différentes manières : elle peut être de type diagnostique ou certificative et avoir lieu à différents moments du cursus (annuelle, à la fin d'un cycle, à la fin de la scolarité obligatoire). Depuis les années 2000, bon nombre de pays ont défini des curricula en termes de compétences, avec des objectifs à atteindre. Les épreuves peuvent être soumises à l'ensemble des élèves d'une cohorte ou à un échantillon, les fonctions pouvant alors différer.

Dans le premier groupe, on trouve un nombre de plus en plus important de pays : la Belgique francophone, le Danemark, la Finlande, la France, la Hongrie, l'Islande, l'Italie, la Lettonie, Malte, la Norvège, les Pays-Bas, la Pologne, le Portugal, le Royaume-Uni (Angleterre, Pays de Galles, Irlande du Nord), l'Écosse ainsi que la Suède.

Dans la plupart des cas, les domaines concernés par l'évaluation sont la langue d'enseignement (L1) et les mathématiques, avec parfois en plus les sciences et la première langue étrangère (L2) (anglais le plus souvent ou allemand, ou autre). Dans certains pays de ce groupe (p. ex. la Finlande), l'évaluation externe peut également porter sur les compétences transversales (p. ex. apprendre à apprendre, motivation à l'apprentissage et communication.) L'évaluation externe se décline de manière assez variable dans ces différents pays. Les épreuves peuvent être obligatoires et passées à tous les élèves en début d'année (de type diagnostique comme p.ex. les évaluations nationales françaises en début de CE2 ou de 6e, voire de CM2 plus récemment) et celles de la Communauté française de Belgique (3e année, 5e, début de 3e année secondaire)[2], ou en fin de 6e année comme aux Pays-Bas. La France

[2] En Communauté française de Belgique, ces évaluations en début d'année remportent un certain succès car elles sont accompagnées d'une analyse et d'un matériel pédagogique pour les enseignants. La France s'est dotée également d'un protocole d'évaluation allant dans le sens d'une analyse qualitative des résultats pour aider les

constitue un exemple intéressant avec deux types d'évaluation externe : l'une ayant pour but d'évaluer les acquis des élèves en début d'année et l'autre, de type évaluation de système (évaluation-bilan), étant passée à un échantillon d'élèves en fin de primaire et de scolarité obligatoire.

La majorité des pays de ce groupe pratiquent une évaluation externe certificative en fin d'année, de cycle, voire de scolarité obligatoire. Un certain nombre de pays comportent une séparation primaire-secondaire I tandis que les pays du nord ont un système organisé de manière continue pour l'ensemble de la scolarité obligatoire. Dans certains cas, les évaluations ont d'abord touché tous les élèves puis se sont adressées à des échantillons d'écoles (p. ex. *INVALSI* en Italie). Le Danemark a introduit récemment des tests nationaux sur ordinateur (d'une durée de 45 mn). Ces tests sont individualisés et adaptés au niveau des élèves qui auront des questions plus ou moins complexes en fonction de leurs réponses.

Au Royaume-Uni (Angleterre, Pays de Galles et Irlande du Nord), il existe des tests nationaux à différents moments du cursus (7, 11 et 14 ans). Les tests ne sont pas utilisés pour mesurer les capacités des élèves mais permettent de comparer les écoles. Pour chacun des domaines considérés, une description en huit niveaux de compétences a été développée.

Dans la plupart de ces pays, ce sont des organismes externes qui s'occupent des tests (*Qualifications and Curriculum Authority* en Angleterre, *INVALSI* en Italie, etc.). En Écosse, les tests nationaux sont considérés comme des moyens d'aider les professeurs dans leurs jugements des élèves par rapport aux standards. Ces derniers sont définis en six niveaux de compétences.

Dans le second groupe, parmi les pays qui ne font pas (encore) appel à des tests nationaux, plusieurs possèdent des curricula définis en termes de compétences, avec des descripteurs de niveaux à atteindre sur lesquels les enseignants peuvent se baser pour déterminer les acquis de leurs élèves (on peut assimiler cela aux atteintes de fin de cycle définies dans les *Objectifs d'apprentissage de l'école primaire genevoise*) : par exemple en Espagne, Roumanie ou encore Chypre. En Allemagne, il n'y a pas de tests nationaux à proprement parler (ils existent dans certains *Länder*, ce qui ressemble à la situation en Suisse) mais un mouvement de définition de standards de formation est en cours comme en Suisse. Cette organisation de l'éducation en Etats ou cantons rejoint celle des États-Unis qui n'avaient pas non plus à proprement parler d'évaluation nationale, même si des études à large échelle ont été réalisées à différents moments (p. ex. NAEP) pour tester les compétences en littératie ou dans d'autres domaines chez les adultes et les jeunes.

Pour ce qui concerne le troisième groupe, les renseignements provenant de la base de données européenne par pays (Eurybase) sont assez peu précis concernant l'évaluation externe.

2. La situation en Suisse romande

Si les épreuves externes existent dans tous les cantons de Suisse romande sauf celui de Berne (partie francophone)[3] (situation en 2007-2008, IRDP, 2007), leurs fonctions et leur fréquence sont toutefois variables. Elles peuvent avoir pour objectif de vérifier l'atteinte des objectifs du programme par les élèves, viser la régulation de l'enseignement ou encore le pilotage du système. Elles peuvent avoir une visée plutôt certificative, d'orientation ou diagnostique. Si dans la majorité des cas, elles ont davantage une fonction sommative ou certificative, visant le contrôle des acquis des élèves, elles peuvent également contribuer au monitorage et à l'évaluation du système, comme le soulignent Wirthner et Ntamakiliro (2008). En effet, dans la plupart des cas, les résultats des élèves sont non seulement diffusés par classe en comparaison avec la moyenne du canton mais également par établissement.

enseignants à établir un diagnostic individuel des points faibles et forts. Il existe également une banque d'outils permettant aux enseignants d'évaluer, quand ils le souhaitent, les différentes compétences de leurs élèves.

[3] En 2007, le canton de Berne avait en 6e des tests communs en français, mathématiques et allemand pour l'ensemble d'une zone de recrutement, élaborés par des enseignants primaires et secondaires et s'adressant aux élèves concernés par la procédure d'orientation.

Dans certains cantons, les finalités des épreuves externes varient selon le degré. Enfin, les domaines évalués au primaire sont le plus souvent le français, les mathématiques et souvent à la fin du primaire l'allemand. Les disciplines sont encore plus variées dans le secondaire.

Ces épreuves sont le plus souvent élaborées par des commissions d'enseignants coordonnées par un responsable des autorités scolaires.

Le tableau 2.1 résume l'état de situation en Suisse romande concernant les épreuves de référence de type certificatif en 2006-2007 décrit par Wirthner et Ntamakiliro.

Tableau 2.1. Épreuves de référence certificatives en Suisse romande en 2006-2007

Degrés	1	2	3	4	5	6	7	8	9
Berne									
Fribourg		●		●		●			
Genève		●		●		●	●	●	●
Jura						●			
Neuchâtel						●			
Valais				●		●		●	●
Vaud		●		●		●		●	

Dans la plupart des cantons, les degrés pour lesquels les épreuves de référence ou épreuves cantonales existent se situent en fin de cycle ou d'école obligatoire (p. ex. 2e, 4e ou 6e, ou encore 9e) ou l'année précédant la fin de la scolarité obligatoire (Genève, Valais, Vaud et bientôt Neuchâtel). Trois cantons ont mis l'accent récemment sur ce type d'épreuves : Genève, Vaud et Valais mais les autres cantons sont également pris dans ce mouvement d'expansion de l'évaluation externe.

Dans certains cantons, on trouve par ailleurs plusieurs types d'épreuves selon les degrés : des épreuves-bilan ou certificatives et des épreuves diagnostiques comme au Jura par exemple en 2e et en 8e. En Valais et à Neuchâtel, des épreuves de référence sont proposées à d'autres moments : par exemple en Valais, épreuves pouvant servir de bilan mais non obligatoires en 2e (français, maths et bientôt environnement), en 3e et en 5e (allemand et environnement ou sciences) ou à Neuchâtel, évaluations sommatives de la 1re à la 3e. Plus récemment, des épreuves-bilan ont été élaborées sur la base du PENSE (nouveau plan d'études cantonal) en français et en mathématiques et seront généralisées en 2008-2009.

Le contenu varie selon la fonction. Les épreuves de fin de cycle avec une fonction d'orientation ou certificative portent sur les objectifs du plan d'études de l'année ou des deux années du cycle, les épreuves ayant une autre visée portent sur un champ plus restreint. Le questionnement peut être également de différents types : privilégier les QCM ou comporter plusieurs formes de questions (fermées et ouvertes). Les épreuves-bilan ont un poids plus ou moins important selon les cantons : par exemple à Neuchâtel, l'épreuve d'orientation de 6e représente un 1/3 de l'évaluation finale pour l'orientation en 7e ; en Valais les épreuves de français et de maths (celle d'allemand n'est pas obligatoire) en 4e, 6e, 8e et 9e représentent 1/5 de la note du deuxième semestre. A Genève, la pondération est variable selon le niveau d'enseignement, primaire ou secondaire : par exemple, en 6P, les épreuves comptent pour 1/3 de la note du troisième trimestre alors qu'au secondaire inférieur, le poids est un peu plus faible (20-25% de la note du troisième trimestre).

Dans le secondaire, la prise en compte des filières varie également selon les cantons. Par exemple à Fribourg en 9e, dans le cadre du diplôme de fin de scolarité obligatoire (français, maths, allemand, anglais ainsi qu'une autre branche à choix), les épreuves sont adaptées aux niveaux des trois types de classes (prégymnasiale, générale et à exigences de base). A Genève (la partie consacrée à Genève sera davantage développée par la suite), on a adopté une autre manière de faire : les épreuves sont en général les mêmes pour les deux types de regroupement A et B (ainsi que pour les classes hétérogènes) mais les barèmes sont différents et adaptés aux deux principaux niveaux.

II. La situation à Genève

Relevons tout d'abord que le nom des épreuves « externes » genevoises diffère entre les deux niveaux d'enseignement : à l'école primaire, il s'agit des épreuves cantonales et au CO, on parle plutôt d'épreuves communes. Comme on a pu le constater dans la situation romande, Genève est le canton qui compte en 2007 le plus grand nombre d'évaluations et de degrés concernés : 2P, 4P et 6P pour le primaire, 7e, 8e et 9e pour le CO. A l'école primaire, les épreuves portent sur le français, les mathématiques et l'écriture-graphisme en 2P puis sur le français, les mathématiques et l'allemand pour la 4P et la 6P. Au CO, le nombre de disciplines n'a fait que croitre depuis quelques années : français, mathématiques, allemand pour les trois degrés, mais également anglais et physique (pour ces deux disciplines, 8e et 9e), biologie (8e), latin (les trois degrés sont concernés). Cette évaluation s'est largement développée depuis quelques années : au primaire, il y a une dizaine d'années, seule l'épreuve de 6P existait (au départ, on parlait d'épreuves d'inspecteurs). Depuis 2000-2001, elles sont organisées à la fin de chaque cycle (2P et 6P) et depuis 2007, également en 4P.

Au CO, les évaluations communes ont été réintroduites de manière massive dans certaines disciplines et en 7e et 8e. L'évaluation commune existe au CO depuis sa fondation. Toutefois, elle a connu des moments plus creux, notamment dans les années 90 (par exemple, en français l'épreuve commune a été suspendue de 1992 à 2001 lors du développement de l'enseignement rénové du français). Au début des années 2000 coexistaient deux types d'épreuves communes au CO : l'évaluation certificative de fin de 9e et l'évaluation diagnostique au début de la 8e (EVADEP), abandonnée car peu exploitée au profit d'une épreuve certificative à la fin de la 7e.

1. Objectifs officiels au primaire et au CO

A l'école primaire, les finalités des épreuves cantonales ont été définies dans un document cadre. « Prioritairement conçues dans la perspective de l'évaluation certificative de fin de cycle les évaluations cantonales contribuent également à la régulation institutionnelle de l'enseignement » (Direction de l'enseignement primaire, document du 12.12.2006). Du point de vue de l'évaluation certificative, elles permettent aux enseignants d'évaluer le niveau de compétences et de connaissances des élèves en fonction des objectifs d'apprentissage. Elles complètent les évaluations effectuées par les enseignants et participent au bilan certificatif de fin de cycle.

Du point de vue de la régulation institutionnelle de l'enseignement, elles interviennent de manière variée. Elles ont pour objectif de réguler les pratiques enseignantes et servent de repères aux enseignants pour situer le niveau de leur classe et de chacun de leurs élèves par rapport à l'ensemble des élèves du canton. Elles jouent également un rôle au niveau de l'école en permettant de construire une représentation collective des attentes du système. Enfin, par rapport aux autorités scolaires, elles permettent d'apporter un éclairage sur les points forts et faibles du système scolaire.

Au cycle d'orientation, l'article 31 du règlement mentionne les objectifs suivants concernant l'évaluation commune :

> « 1) L'évaluation commune a pour but de permettre à l'élève, à ses parents et à l'école de situer l'élève par rapport à la maîtrise des notions et des compétences acquises au moment où les épreuves ou les tests sont administrés et par rapport aux autres élèves.
>
> » 2) L'évaluation commune peut être formative et permet d'ajuster le travail de l'élève et de la classe à partir des observations faites par la maitresse ou le maitre.

> » 3) L'évaluation commune peut être certificative dans la mesure où, à la fin d'une étape inscrite au plan d'études, elle valide les acquis des élèves.[4] »

Plus récemment, des cadres de travail concernant l'élaboration des EVACOM[5] ont été conçus dans un certain nombre de disciplines. Outre les nombreuses indications se rapportant à la conception technique des épreuves, ils en précisent les buts dans les termes suivants :

« Les évaluations communes contrôlent les atteintes des principaux objectifs du français (par exemple) par les élèves des différents degrés. Elles certifient les acquis des élèves. Elles s'inscrivent dans une perspective de régulation de l'enseignement. »

De manière générale, aussi bien à l'école primaire qu'au CO, les épreuves cantonales ou communes ont plusieurs fonctions : une fonction certificative et une fonction de régulation. On peut se demander si dans les deux niveaux d'enseignement, cette deuxième fonction recouvre exactement la même chose : régulation *institutionnelle* de l'enseignement *vs* régulation de l'enseignement. L'école primaire est à ce sujet plus explicite concernant les différents niveaux sur lesquels la régulation doit porter (enseignant, école, autorités scolaires).

2. Le point de vue des autorités, des directions de l'enseignement[6] et des concepteurs des épreuves cantonales et communes[7]

2.1. Représentations des objectifs par les différents acteurs (autorités, des directions de l'enseignement et des concepteurs des épreuves cantonales et communes)

Pour les deux membres du secrétariat général (secrétaire général et secrétaire adjoint), les évaluations cantonales et communes peuvent avoir plusieurs fonctions, qu'il s'agirait de préciser dans le cadre de ce projet. De manière générale, l'évaluation cantonale ou commune permet selon eux d'obtenir des informations et vérifier les compétences et les connaissances des élèves en regard du plan d'études (fonction de certification des élèves), renseigner au niveau du pilotage du système (fonction de pilotage du système au niveau des directions générales et des établissements), réguler les pratiques d'enseignement et d'évaluation (fonction de régulation de l'enseignement), notamment dans les groupes de disciplines. Ces différentes dimensions ont pour effet d'harmoniser le système genevois. Ces objectifs se concrétisent un peu différemment entre le primaire et le CO : à l'école primaire, on chercherait à vérifier que la majorité des élèves maitrisent les compétences de base ou à savoir combien d'élèves les ont acquises en lien avec le programme et les objectifs d'apprentissage tandis qu'au CO, on chercherait davantage à situer les élèves les uns par rapport aux autres et à regarder à quel niveau ils sont dans le processus d'acquisition. Ces deux focalisations quelque peu différentes ont inévitablement un effet sur le seuil d'exigences de l'épreuve. Par ailleurs, on peut observer des différences entre disciplines.

---

[4] D'autres indications sont également données sur les modalités et l'élaboration des épreuves :

> « 4) Les tests et les épreuves communes sont administrés à l'ensemble des élèves d'un degré. Le calendrier est communiqué aux élèves et à leurs parents au début de l'année scolaire. 5) L'évaluation commune est conçue sous la responsabilité de la direction générale. Les travaux sont élaborés par des groupes de travail relevant des présidences des groupes des disciplines concernées, en étroite collaboration avec le service de l'évaluation commune. »

[5] Le sigle EVACOM désigne deux éléments : d'une part, les épreuves communes du CO et d'autre part, l'application informatique élaborée pour saisir les données puis définir les barèmes et donner les résultats.

[6] Au CO, la directrice de l'enseignement interviewée est Mme B. Badoud-Volta qui assurait cette tâche jusqu'en novembre 2008.

[7] Dans cette partie, les propos sont ceux de M. F. Bugniet, responsable de l'évaluation commune jusqu'en juin 2008.

Au niveau des directions de l'enseignement, la principale fonction des évaluations cantonales ou communes est de (contribuer à) certifier les élèves en conformité avec les plans d'études. Étant donné qu'elles ne comptent que pour une partie de la note du troisième trimestre (1/3 pour le primaire[8] en 4P et 6P ; 1/5 à 1/4 au CO selon les disciplines), elles participent à cette certification. Toutefois, elles ont un poids symbolique plus élevé que leur valeur effective et figurent séparément dans le livret scolaire. A l'intérieur d'un ordre d'enseignement, l'accent est un peu différent selon les degrés : par exemple, en 2P *vs* 6P ou en 7e ou 8e *vs* 9e, degré pour lequel la pression est plus forte. D'ailleurs, les écoles et associations professionnelles voudraient utiliser ces épreuves comme examen d'entrée[9].

Les évaluations cantonales ou communes jouent également un rôle de régulation de l'enseignement en montrant les attentes institutionnelles.

En revanche, pour la directrice de l'enseignement du CO, il est difficile de leur faire également jouer un rôle de régulation du système, puisqu'elles ne sont pas faites pour ça.

Les deux coordinateurs des épreuves (l'adjoint à la direction de l'enseignement primaire et le responsable du secteur de l'évaluation commune du CO) ont des représentations assez proches : au primaire, la principale fonction est la certification. Au CO, il s'y ajouterait une fonction de régulation de l'enseignement. Le responsable souligne également l'accent différent selon les degrés et le statut particulier conféré aux épreuves communes de 9e. Enfin, le responsable de la préorientation relève le fait que les épreuves étant communes à tous les élèves et diffusées à tous les enseignants, elles ont pour effet d'inciter les enseignants à aborder les notions, les domaines qui sont dans ces épreuves communes.

Au niveau des concepteurs d'épreuves, on observe des différences de représentations aussi bien au niveau des disciplines prises en compte que du niveau d'enseignement.

A l'école primaire, en français, les épreuves cantonales ont pour objectif de faire passer un enseignement-apprentissage dans les classes. Elles ont une fonction de levier, d'introduction (p. ex. de la production écrite) ou de changement de certaines pratiques, même si ce n'est pas leur rôle premier. Elles permettent de montrer ce qu'on peut attendre des élèves et définissent une progression des domaines langagiers en termes d'attentes institutionnelles. En allemand, elles ont pour principal objectif d'attester d'une compréhension orale et écrite de la langue allemande. En mathématiques, les concepteurs leur voient une double fonction (reprise des textes officiels) : l'évaluation des acquis des élèves (prioritairement) et la régulation du système. Enfin, au secteur de l'évaluation et de la différenciation, on souligne l'existence de plusieurs fonctions, ce qui peut poser problème : un rôle certificatif (atteinte des objectifs par les élèves), un moyen de réguler et également de servir au pilotage de l'institution grâce à « des analyses de plus en plus fines » (comparaison entre établissements, entre classes à l'intérieur des écoles, etc.).

Au *cycle d'orientation*, on observe là aussi des différences de conceptions selon les disciplines. En français, où l'épreuve commune a été suspendue pendant de nombreuses années et n'a pas été réintroduite avec un grand enthousiasme, on relève que l'objectif principal est de vérifier que le plan d'études est suivi et accompli dans toutes les classes. Il est également souligné une tension entre deux tendances : un bilan de compétences des élèves en fin de 9e (une sorte d'examen à la fin de la scolarité obligatoire) mais aussi un outil de régulation de l'enseignement, ce que les commissaires des épreuves privilégient.

En allemand, c'est la certification qui est avant tout mise en avant. En mathématiques, les commissaires mettent en évidence le fait qu'une fois par an, on bénéficie d'une évaluation « standardisée » commune à tous les élèves d'une cohorte et qui permet de vérifier si les enseignants ont réalisé le programme. Elle permet également de situer les élèves les uns par rapport aux autres.

[8] En 2P, étant donné qu'il n'y a pas de notes, elles entrent dans le bilan certificatif.

[9] Il faut préciser qu'il y a quelques années, les écoles professionnelles faisaient passer leur propre examen d'entrée aux élèves sortant du CO, ne se contentant pas des indications figurant dans leur livret scolaire. Actuellement, ces écoles utilisent la partie « tronc commun » des épreuves communes de 9e dans certains domaines comme le français ou les mathématiques.

Elle donne lieu à une note. Il s'agit également d'une épreuve certificative qui fixe un certain standard, une norme de réussite et permet de regarder quels points du plan d'études ont été acquis. Ils soulignent également la place particulière des épreuves en mathématiques et français de 9e dont le tronc commun est utilisé pour l'entrée en apprentissage. Enfin, en physique, elle permet une autorégulation, c'est-à-dire de savoir si les attentes que l'on se fixe sont raisonnables ou non et de se remettre en question par rapport à ce qu'on fait.

On pourrait faire l'hypothèse que ces différentes centrations auront un impact sur la conception, le contenu et le seuil d'exigences des épreuves.

2.2. Déroulement de l'élaboration des épreuves

Organisation de l'élaboration des épreuves cantonales et communes

L'organisation de l'élaboration des épreuves cantonales et communes varie selon l'ordre d'enseignement concerné. A *l'école primaire,* sous la responsabilité de la directrice de l'enseignement et de l'adjoint à la direction qui coordonne les différentes épreuves cantonales, des formateurs de l'enseignement primaire (didacticiens des disciplines en collaboration avec les formateurs du secteur de l'évaluation et de la différenciation) conçoivent les épreuves des trois disciplines (français, mathématiques et allemand[10]) pour les trois degrés considérés (2, 4 et 6P). Au *cycle d'orientation*, ce sont des commissaires qui réalisent les épreuves communes dans les nombreuses disciplines soit pour l'ensemble des degrés, soit pour seulement un, voire deux degrés selon les cas, sous la supervision et la coordination des présidents de groupe des différentes disciplines. Les commissaires sont des enseignants qui sont dégrevés pour quelques heures par semaine (1h30-2 heures par personne soit une demi-journée au total) pour effectuer ce travail. Ils travaillent le plus souvent à deux par degré.

Les deux coordinateurs des épreuves cantonales et communes (école primaire et CO) participent aux différentes phases mais sont également impliqués dans les aspects administratifs et de gestion des épreuves cantonales et communes ainsi que dans la phase d'analyse et de diffusion des résultats auprès des enseignants et des directeurs d'établissement ou inspecteurs.

A *l'école primaire,* le déroulement de l'élaboration des épreuves comporte quatre étapes :

1) en mai-juin, les responsables de la discipline en question réalisent un avant-projet (texte ou thème choisi) qu'ils soumettent à la directrice de l'enseignement et au coordinateur des épreuves pour avis. Ces derniers valident (ou non) l'avant-projet.

2) Les concepteurs-formateurs du CEFEP (Centre de formation de l'enseignement primaire) élaborent l'épreuve et la font valider au coordinateur.

3) L'épreuve est prétestée en début d'année scolaire dans trois classes avec composition sociodémographique contrastée. L'analyse du prétest et des questions est également présentée à la directrice de l'enseignement. En fonction des résultats et de cette consultation, l'épreuve est modifiée, les items trop faciles ou trop difficiles sont éliminés.

4) Une fois ces modifications effectuées, l'épreuve est soumise pour lecture et validation institutionnelle finale en novembre-décembre à la directrice de l'enseignement.

Au *cycle d'orientation*, la situation est quelque peu différente et varie également selon les disciplines. Les commissions d'enseignants déterminent le champ qu'elles soumettent au président de groupe de la discipline concernée pour validation. Le responsable du secteur de l'évaluation commune (SEC) procède ensuite à une autre validation par rapport au champ attendu selon les objectifs du plan d'études. Il peut aussi demander s'il y a une cohésion au sein du groupe de disciplines (composé des responsables de disciplines des vingt établissements). Les commissaires élaborent ensuite les épreuves par degré et le président de groupe relit l'épreuve, ainsi que des membres du conseil de direction et

[10] En 2P, une épreuve porte également sur l'écriture-graphisme. Il n'en sera pas question dans cette étude.

une troisième personne spécialiste en évaluation ou de la discipline (situation jusqu'à 2007-2008). Le responsable du SEC effectue une dernière lecture et essaie les différents exercices ainsi que les critères de correction. La directrice de l'enseignement participe également à cette dernière lecture (situation jusqu'à 2007-2008). Les épreuves doivent être prêtes au plus tard en mars.

Les prétests sont variables selon les disciplines, comme on pourra le constater[11].

Le tableau 2.2 résume le processus d'élaboration des épreuves dans les deux ordres d'enseignement.

Tableau 2.2. Organisation et déroulement de l'élaboration des épreuves cantonales ou communes (EC) dans les deux ordres d'enseignement

	École primaire	CO
Qui ?	Coordinateur EC (expert en docimologie) Concepteurs (formateurs du CEFEP : didacticiens des différentes disciplines + experts en évaluation)	Coordinateur évaluation commune Commissaires d'épreuves (enseignants) par discipline et degré sous la responsabilité du ou des PG (dégrèvement)
Comment ?	*4 étapes :* 1) en mai-juin : avant-projet réalisé par les concepteurs (didacticiens des disciplines) soumis à la directrice de l'enseignement et au coordinateur pour validation 2) Élaboration de l'épreuve par les concepteurs didacticiens, validation par le coordinateur 3) Début d'année scolaire : prétest dans trois classes ; analyse du prétest présentée à la directrice Modifications en fonction du prétest et des remarques de la directrice 4) Novembre-décembre : épreuve soumise pour lecture et validation institutionnelle finale	*5 étapes :* 1) Champ de l'épreuve déterminé par commissions d'enseignants et soumis pour validation au PG 2) Validation par le coordinateur, responsable du SEC en fonction du champ attendu selon le plan d'études 3) Élaboration des épreuves par les commissaires et relecture par le PG puis par des membres du conseil de direction et d'autres personnes spécialistes en évaluation ou de la discipline concernée 4) Relecture par le responsable du SEC qui essaie les exercices et teste les critères de correction 5) Jusqu'en 2008, dernière relecture par la directrice de l'enseignement Délai final : mars [Dans certaines disciplines, il y a également des prétests mais pas de façon systématique]

Directives et indications concernant l'élaboration des épreuves

Les directives et indications sont de différentes natures et revêtent différents statuts. A *l'école primaire*, la direction a rédigé des documents définissant les objectifs des épreuves cantonales, le cadre, le calendrier ainsi que les responsabilités des différents intervenants. Au secteur de la différenciation et de l'évaluation (SEDEV), on nous a également montré un document fournissant des indications plus « docimologiques » qui n'est pas véritablement officiel. Le contenu est déterminé par les « Objectifs d'apprentissage à l'école primaire » où figurent des attentes de fin de cycle.

Au *cycle d'orientation*, il y avait jusqu'en 2008 peu d'indications sur l'élaboration des épreuves. Excepté les fonctions, le calendrier et la répartition des tâches, les informations étaient tacites (p. ex. règles concernant les questions : éviter les QCM avec trois propositions et les distracteurs trop évidents). Depuis 2007, les procédures ont été formalisées dans certaines disciplines comme le français et les mathématiques. Les prétests ont été introduits petit à petit : la situation est variable et dépend de la planification de l'épreuve. Dans certaines disciplines, les commissaires sont parvenus à avoir une année d'avance (situation jusqu'à 2007-2008). Le contenu est dépendant du plan d'études et

[11] L'objectif fixé par la direction du CO est d'avoir une année d'avance pour toutes les épreuves et de généraliser les prétests.

on tient compte du poids relatif des différents objectifs et domaines dont la proportion est précisée de manière variable selon les disciplines.

Selon la discipline, on gère différemment la prise en compte des regroupements A et B et des classes hétérogènes. Dans certains cas, l'épreuve de B est incluse dans celle de A, dans d'autres cas il s'agit de la même épreuve et on applique des barèmes différents (comme en français).

Élaboration des épreuves selon les différentes disciplines

A l'école primaire

En français, les concepteurs déclarent qu'au départ ils disposaient de peu d'indications. Ils relèvent certaines divergences entre leur secteur et celui du secteur différenciation-évaluation des apprentissages (SEDEV). Les représentants de ce dernier souhaiteraient qu'il y ait le même nombre de questions selon les différentes composantes de l'épreuve, ce qui parait difficile aux concepteurs selon les domaines (possible pour la structuration, difficile voire peu pertinent pour la compréhension et la production écrite, car on se situe davantage dans des compétences complexes où toutes les composantes n'ont pas véritablement la même importance). Il est aussi difficile d'évaluer de la même manière les différentes composantes et de déterminer un seuil de réussite pour chacune d'entre elles, notamment en production écrite. A cela, s'ajoutent les théories didactiques sous-jacentes et plus particulièrement les genres textuels. Pour les didacticiens en charge des épreuves, quand un élève comprend un texte d'un genre spécifique, cela ne signifie pas qu'il est capable de lire des textes de tous les genres.

Le contenu varie d'une année à l'autre, en particulier le choix des domaines du français I. Seule la compréhension de l'écrit est présente pour chaque degré et chaque année, la compréhension de l'oral et la production écrite sont prises en compte de manière variable selon les années. Il est communiqué aux enseignants en début d'année, mais pas le genre textuel comme c'était le cas il y a quelques années. Les enseignants ont à leur disposition une planification qui prévoit de travailler trois genres textuels définis par an.

L'élaboration de l'épreuve repose sur le choix du texte puis la conception des questions. Les pré-tests sont appréciés car ils donnent des informations sur la difficulté des questions, notamment au niveau de la formulation et de la compréhension par les élèves. Ils permettent également d'éliminer les questions trop faciles ou trop difficiles.

Les épreuves ont également évolué au niveau de la mise en contexte qui était assez longue au départ et s'est atténuée suite aux indications du coordinateur pour éviter les biais et les différences au niveau des conditions de passation.

En allemand, le contenu porte sur la compréhension orale et écrite ainsi que sur l'expression écrite. Pour l'instant (jusqu'en 2007-2008), l'expression orale n'est pas prise en compte car elle prendrait beaucoup trop de temps. Là aussi, les épreuves ont évolué depuis 1999. L'introduction de la méthode *Tamburin* ainsi que le cadre européen des langues ont changé les pratiques. Les épreuves sont calquées sur la méthode. Il est souligné que le niveau d'exigence a également évolué : d'après la responsable, il a baissé pour la fin de la 6P puisque l'objectif attendu est aujourd'hui le niveau A1. Elle regrette également une certaine homogénéisation des questions (beaucoup de QCM, voire des questions fermées à réponse courte) et donc la perte en diversité, la méthode utilisée ne permettant plus vraiment les questions portant sur l'implicite.

Le champ testé étant basé sur la méthode, il n'est communiqué ni aux enseignants ni aux élèves. Les trois personnes impliquées travaillent de manière transversale.

En mathématiques, les contenus diffèrent selon les années. En 2P, il y a moins d'objectifs d'apprentissage et donc moins d'items possibles (8-9 items). Les concepteurs ont défini des objectifs incontournables et pour le reste un tournus est établi. Pour les deux autres degrés, le choix est moins difficile mais certains aspects ne peuvent pas être testés par des épreuves papier-crayon. De manière générale, les formateurs proposent des items. Tous les volets du plan d'études sont pris en compte. Certaines règles leur ont été données par la direction de l'enseignement : 1/3 consacré au domaine de

l'espace et 2/3 à celui du nombre. Par ailleurs, les questions doivent être pour 1/3 de type application et vérification d'outils ou de connaissances et de 2/3 de type situation-problème.

Les formateurs observent une certaine standardisation notamment en 2P où la flexibilité est moins grande, les items supposant une interaction maitre-élèves n'étant pas les bienvenus. Or, dans les classes, ce type de pratique est courant.

Les items sont testés séparément pendant l'année, parfois sous plusieurs formes avant le prétest de l'ensemble de l'épreuve dans les trois classes. Les épreuves sont très comparables d'une année à l'autre comme on peut le constater dans les tables de spécification[12]. Les formateurs ont constaté que les épreuves sont de plus en plus difficiles (notamment en 2P). Toutefois, le taux de réussite reste stable, les élèves s'étant accoutumés à l'existence d'épreuves.

Enfin, la formatrice en évaluation dont le rôle est d'intervenir une fois qu'un premier jet de l'épreuve a été réalisé, déclare essayer de sensibiliser un peu la commission quant aux différentes manières d'évaluer, sur les différents types d'exercices. Leur rôle consiste à vérifier la table de spécification, notamment la mise de points et la correspondance par rapport aux objectifs visés. Elle souligne combien il est difficile d'estimer le niveau de difficulté des items, d'où la nécessité des prétests et des essais de questions présentées de plusieurs manières. Elle relève également l'évolution de ces épreuves dont le niveau d'exigences a augmenté, sans doute sous la pression de certains groupes (ARLE, par exemple) : les questions doivent discriminer, les questions trop faciles sont éliminées.

Au cycle d'orientation

En français, depuis 2007, sous l'impulsion de la direction de l'enseignement, un cadre a été rédigé par L. Weiss et les présidents de groupe pour rendre les épreuves plus homogènes. Auparavant, les acteurs impliqués s'inspiraient du guide méthodologique (Pini, Reith, Weiss et Bugniet, 2006).

Les commissaires soulignent les difficultés à élaborer une épreuve (notamment en 7^{e}) qui s'adresse non seulement à des élèves des regroupements A et B ou de classes hétérogènes, mais également à des élèves du regroupement C[13]. Pour eux, l'épreuve de français qui porte sur un plan d'études très copieux prend plus de temps à élaborer que celle d'autres domaines (p. ex. les maths) car elle n'est pas transposable d'une année à l'autre. En effet, elle repose tout d'abord sur un long travail de choix du texte. Ce choix est rendu encore plus difficile pour la 9^{e} où l'on teste l'ensemble du programme. Trouver le bon texte est considéré comme un apprentissage en soi et prend du temps. Une fois le texte trouvé sur la base des types de textes du degré considéré, on élabore des questions en fonction de grands chapitres incontournables. Une partie des questions sont stables (notamment pour la composante « langue », synonymes, antonymes, etc.). Les épreuves sont prétestées dans une ou deux classes, surtout pour tester les formulations des questions. Cette année, le prétest sera effectué dans un autre canton (Neuchâtel).

Depuis quelques années, la priorité est donnée à la lecture, ce qui oriente le type de compétences évaluées. Par ailleurs, la contrainte liée à la réalisation d'une épreuve unique pour tous les regroupement incite à axer les questions sur des savoir-faire vraiment fondamentaux. Il s'agit à la fois de détecter l'excellence chez certains élèves et les connaissances acquises même si elles ne sont pas complètement formalisées chez les élèves faibles. Enfin, un choix a été fait par le groupe de choisir un thème pour l'épreuve.

En allemand, depuis 2 ans, les commissaires doivent effectuer pour chaque degré deux ensembles d'épreuves en fonction des deux méthodes coexistant, les unes liées à *Sowieso* et les autres à *Genial* (méthode introduite dans trois établissements et à l'école primaire). Les épreuves sont également adaptées aux regroupements et au niveau des élèves. Depuis l'introduction du cadre européen des

[12] Les tables de spécification dont on trouvera un exemple dans l'annexe (2) précisent pour chaque domaine évalué dans l'épreuve l'objectif visé et le nombre de points attribués.

[13] Seuls les deux commissaires en charge de l'épreuve de français de 7^{e} mentionnent le regroupement C. Il faut préciser que ce regroupement n'existe pas dans tous les collèges ni dans tous les degrés. Cet élément est probablement particulièrement crucial dans le cas d'une épreuve de 7^{e} dans une discipline sans niveaux différenciés.

langues (et du Portfolio européen des langues, PEL), l'évaluation a changé et a nécessité de se conformer au cadre.

En 2007-2008, le groupe a repris et adapté pour la 9e une épreuve du canton de Fribourg et a testé dans trois écoles l'évaluation de l'expression orale. De manière générale, l'orientation porte sur les aspects communicatifs et les épreuves comportent de la compréhension orale, de la compréhension et de l'expression écrites. La grammaire a été plutôt mise de côté même si elle intervient de plus en plus dans les différents domaines au fil des degrés : chacun de ces domaines est également représenté dans l'épreuve (1/3 par domaine). On essaie également de varier le type de questions mais il n'y a pas de question ouverte. La durée est différente selon le regroupement (2h pour les A et les H niveau fort et 45 mn pour les B). En raison du temps et de la multiplication des épreuves[14], le groupe n'a pu prétester les épreuves et a réalisé quelques essais individuels et ponctuels.

En mathématiques, les épreuves communes portent généralement sur 2/3 voire 3/4 du programme. Jusqu'en 2007-2008, le groupe élaborait les épreuves sur la base d'instructions tacites. Maintenant, un cadre a été rédigé comme pour le français. Le champ est défini par le groupe de disciplines composé des responsables de disciplines (RD) des vingt établissements pour les trois degrés. L'option prise est de tester les objectifs du plan d'études dans les différents domaines. Une partie ne change pas d'une année à l'autre et concerne les connaissances essentielles pour lesquelles le groupe a travaillé pendant deux ans et a essayé de fixer le niveau minimum à atteindre pour les deux regroupements A et B, et ce pour l'ensemble du plan d'études. Les commissaires répartissent les questions en problèmes ouverts et exercices de drill. Toutefois, afin de standardiser le plus possible la correction, les problèmes ouverts sont minoritaires. Les épreuves durent 2h. Étant donné que les commissaires ont réussi à élaborer leurs épreuves avec une année d'avance, ils les prétestent en avril/mai. Le champ de l'épreuve est diffusé de manière globale en début d'année scolaire puis de manière plus détaillée à la fin du 1er semestre.

En physique, le mandat sur l'évaluation confié par le directeur général du CO a permis de clarifier les choses : le contenu du plan d'études a été classé en termes de compétences globales et d'objectifs plus restreints, comme c'est le cas dans le Plan d'études romand actuellement en consultation. Pour les trois cours de physique, les attentes ont été validées par les RD. On alterne les sujets dans les épreuves. Une autre particularité de ce domaine consiste à déterminer deux niveaux de difficultés, ce qui permet de prévoir, contrairement aux autres disciplines, un seuil de réussite[15]. Jusqu'en 2007-2008, il était très difficile de prétester les épreuves ; cela sera maintenant possible, étant donné que les concepteurs ont de l'avance.

Le déroulement de la conception de l'épreuve est le suivant : chaque commissaire peut proposer une ébauche de l'épreuve avec des questions qui font ensuite l'objet d'une discussion et d'un travail sur leur contenu et sur les critères de correction.

2.3. Modalités : passation, correction, barèmes

A l'école primaire comme au CO, la passation et la correction des épreuves sont assurées par les enseignants titulaires des élèves ou ceux dispensant la discipline concernée.

De manière générale, la façon de déterminer les barèmes varie également entre les deux niveaux d'enseignement. A l'école primaire, les barèmes sont déterminés a priori (sur la base des essais, tout de même), le seuil de réussite étant fixé à 2/3 des points. Le barème est critérié et repose sur une table de spécification établie en fonction des objectifs visés et des différentes composantes d'un domaine donné. Au CO, la situation est un peu plus variable. Dans la plupart des cas, les barèmes sont déterminés a posteriori sur la base de la distribution des résultats des élèves et adaptés aux différents regroupements (A, B, H). Parfois, comme en physique ou en latin, ils sont définis a priori et adaptés si

[14] C'est la dernière année que les deux méthodes coexisteront.

[15] Il ne s'agit pas d'un seuil a priori mais on s'en rapproche. Par ailleurs, cette estimation concorde bien avec les résultats des élèves.

besoin en fonction des résultats obtenus par les élèves. Pour parvenir à déterminer des seuils par objectifs, il est nécessaire d'avoir une vision relativement précise de ce qu'on attend des élèves, d'avoir stabilisé une forme d'épreuve. Le responsable du SEC précise qu'en physique, par exemple, les PG et les commissaires n'avaient pas regardé les résultats mais avaient estimé la difficulté de l'épreuve et cela correspondait aux résultats des élèves.

A *l'école primaire*, malgré une certaine coordination, on observe quelques différences d'interprétation selon la discipline. En français, les concepteurs ont renoncé à des seuils de réussite par composante. Par exemple en production écrite, ils élaborent une grille de critères et essaient de voir ce qu'un texte doit montrer ou contenir au mieux. Le seuil de réussite et le barème sont fixés à la suite des prétests, mais il subsiste selon eux une part d'inconnues. Le groupe de concepteurs relève également les pressions des autorités ou des politiques qu'ils ressentent. Parfois, les directives iront dans le sens de rendre plus faciles les épreuves, ou au contraire plus difficiles. La question de la structuration en 2P qui jusqu'en 2008 n'était pas certificative mais formative et ne figurait pas dans les résultats est un exemple également intéressant puisque son statut a changé (elle y figure actuellement). On a pu aussi constater des différences de réussite entre compréhension et production écrites. La consigne donnée par la direction de l'enseignement primaire était de durcir les exigences pour la compréhension de l'écrit.

En allemand, les attentes sont exprimées au niveau de chaque item. Si l'item est facile, on attend une réussite élevée (seuil à 2 points sur 3), s'il est difficile, le seuil d'exigences sera plus bas. L'épreuve est calibrée sur 60 points avec une vingtaine de points pour la compréhension orale, la compréhension écrite et 10 points pour l'expression écrite (qui comporte une dimension structuration).

En mathématiques, le groupe de concepteurs souligne le fait qu'il a peu de liberté, le seuil de réussite est déterminé à 2/3 des points. On leur conseille de mettre des codes partiels quand les questions ont 2 points par exemple. Ils ne font plus figurer de seuil de réussite par question dans la table de spécification mais des seuils par domaine (espace, nombre) ou par type d'exercice (situation-problème ou application).

Le secteur de l'évaluation et de la différenciation qui a été formé spécifiquement à l'évaluation par L. Allal, anciennement professeur à la FAPSE sur la problématique de l'évaluation, relève la nécessité d'avoir certaines règles : le seuil de réussite doit être bien équilibré (un point par item) ; si un objectif est à travailler, il faut multiplier le nombre d'items ; en cas de questions à 2 points (par exemple en maths), il faut réfléchir où placer le seuil de réussite.

Leur rôle est de veiller à ce que l'épreuve cantonale soit équilibrée par rapport aux différentes composantes de l'épreuve même si tout le monde ne s'accorde pas sur la difficulté d'un exercice.

Au *cycle d'orientation*, une question cruciale revient chez les différents interlocuteurs : l'adaptation du barème aux élèves de B.

En français, le seuil de réussite est déterminé a posteriori sur la base de la moyenne et de la distribution des résultats. Il n'y a pas de table de spécification. Les résultats sont reportés par partie de l'épreuve en fonction de trois composantes (contenu, moyens langagiers et langue). Les commissaires sont relativement critiques par rapport à la table de spécification telle qu'elle est établie au primaire et ne partagent pas l'approche de leurs collègues où tout est fixé par avance et est adapté par la suite sur la base des prétests. Pour eux, les concepteurs du primaire partent des objectifs du plan d'études (complexes) qu'ils testent par une question à 3 pts (par exemple) et ensuite ils évaluent la question (acquis à 2 pts, non acquis à 1 pt). La question qui revient est de savoir si l'épreuve sert à faire un bilan de compétence ou est un outil de régulation.

En allemand, on adapte le barème : en A, 75% des élèves réussissent le 66% de l'épreuve. Cela est nettement plus compliqué lorsqu'il s'agit de fixer un barème pour les élèves de B. Ces derniers ont non seulement des problèmes liés à la lecture mais aussi en compréhension de l'oral (c'est pourquoi ils ont introduit une troisième écoute du texte à comprendre).

En mathématiques, il n'y a pas non plus de barème a priori. Le seuil de réussite (3.5) tourne autour de 60% pour les A et 50% pour les B. Le groupe relève également les difficultés à fixer le barème pour les B (niveau normal) dont les résultats sont souvent inférieurs au 50% des points de l'épreuve. De

plus, le tronc commun de l'épreuve de 9e où l'on teste beaucoup le programme de 8e est d'abord élaboré pour les A puis adapté pour les B.

Enfin, en physique, comme on a déjà pu le constater, les seuils de réussite sont estimés grâce aux objectifs et aux deux niveaux de difficulté des items. Le groupe n'éprouve pas le besoin d'établir le barème sur la base de la distribution grâce aux niveaux de difficulté fixés au préalable. Les items sont classés dans deux niveaux de difficulté : niveau 1 (items qui portent sur des connaissances) et niveau 2 (items plus complexes). Quant aux commissaires et aux PG, ils attendent des élèves de A, 80% de réponses justes aux items de niveau 1 et 60% à ceux de niveau 2. Pour les élèves de B, les proportions tombent respectivement à 60 et 40%. Ces exigences marchent bien en classe mais sont à réduire de 5% pour les épreuves communes (stress et formulation inédite des questions pour les élèves). Les épreuves devraient idéalement être composées de 2/3 de questions de niveau 1 et de 1/3 de niveau 2, mais parfois certaines questions s'avèrent plus difficiles que prévu.

2.4. Analyse des résultats des élèves aux épreuves cantonales et communes

Au *primaire*, les résultats reportés sur une grille par les enseignants sont transmis à la direction de l'enseignement primaire qui se charge de leur saisie informatique puis les analyse et renvoie les résultats à chaque école et chaque enseignant. Les informations transmises portent sur le score des élèves, la moyenne du canton, et la proportion d'élèves ayant atteint le seuil de réussite (taux de réussite). Depuis trois ans, une analyse est également réalisée par école et diffusée à la rentrée, permettant de comparer les résultats de l'école à l'ensemble du canton.

Au *cycle d'orientation*, la situation est différente. Il existe une application informatique très élaborée, EVACOM, qui permet aux enseignants d'entrer les résultats de leurs élèves. Le degré de précision dépend des indications fournies par les responsables des épreuves communes dans la discipline. Il y a des différences entre disciplines : par exemple en mathématiques, on entre également les résultats des élèves par question, ce qui permet une analyse par item. Pour le responsable du SEC, si l'outil est performant, il reste encore à améliorer : pour l'instant il est difficile de faire des modifications (par exemple, enlever une question si elle est trop mal réussie par l'ensemble des élèves). Les analyses et le barème peuvent être réalisés rapidement dès que tous les enseignants ont saisi les résultats de leurs élèves : histogramme avec fréquences, fréquences cumulées, rendement, nombre de points. Les barèmes sont différenciés en fonction des regroupements (A et B) et des types de classes (classes hétérogènes). Dans certains cas comme l'anglais, cela peut poser problème avec les classes hétérogènes, étant donné que contrairement à l'allemand et aux maths, il n'y a pas de niveaux et ceci malgré la présence d'élèves que l'on pourrait assimiler à des élèves de regroupement B. Les enseignants reçoivent les résultats de leur classe et peuvent les comparer à ceux de l'ensemble des élèves du regroupement ou des classes hétérogènes.

De manière générale, en plus des analyses diffusées aux enseignants et aux établissements, il n'y a pas véritablement de retour au niveau des groupes de concepteurs. A l'école primaire, les formateurs déclarent qu'ils reçoivent une information générale qui porte sur les résultats globaux de l'ensemble du canton, mais ni par école, ni par question, ce qui leur serait certainement utile. Au CO, malgré l'existence d'EVACOM, il y a peu de temps à disposition pour véritablement tirer profit des résultats.

2.5. Utilisation des épreuves

De par l'organisation des deux niveaux d'enseignement, les utilisations sont assez différentes. En effet, à l'école primaire, ce sont les formateurs qui conçoivent les épreuves. Ils peuvent les utiliser lors de formations continues dans les écoles. Au CO, ce type d'utilisation n'est pas possible étant donné que les présidents de groupe organisent les formations mais ne les dispensent pas. Ils peuvent au mieux intervenir au niveau de leur établissement (discussion avec les RD).

Au niveau des deux directions, on se pose des questions un peu différentes. A *l'école primaire*, une exploitation plus ciblée des résultats au niveau du terrain des écoles est prévue, notamment pour définir les objectifs prioritaires des projets d'établissement. Il serait intéressant de voir l'évolution sur 4 ans et par exemple les décalages entre domaines. La directrice de l'enseignement souligne également la nécessité de bénéficier des gens du terrain pour effectuer un travail d'accompagnement dans les écoles. Elle relève l'importance de disposer d'une base de données par établissement.

Au *cycle d'orientation*, l'évaluation commune donne un éclairage sur l'orientation quand on lit ses résultats à la lumière de ceux des autres évaluations. La directrice d'enseignement relève la faible pondération de cette évaluation contrastant avec une valeur symbolique forte. Par ailleurs, les évaluations communes sont utilisées dans les classes à titre d'entrainement, si l'on en juge par le nombre d'épreuves envoyées par le responsable du SEC aux enseignants, parents voire écoles privées.

Du côté des concepteurs du *primaire*, les épreuves ou plutôt les supports (textes en français ou en allemand par exemple) sont utilisés dans la formation ou dans des séquences didactiques. Les supports créés dans le cadre des épreuves (par exemple, des interviews) peuvent être assimilés à des modèles du genre que les enseignants pourraient utiliser également comme moyens d'enseignement. En mathématiques, les concepteurs relèvent que certaines écoles font appel à eux pour faire une formation quand les résultats de leurs élèves sont bas. Au secteur de la différenciation et de l'évaluation, les épreuves cantonales sont moins utilisées en formation que lors du passage du certificatif au formatif, notamment en travaillant sur les tables de spécifications (objectifs, attribution de points, seuil de réussite, type d'exercices).

Au *cycle d'orientation*, il y a peu de réutilisation des épreuves si ce n'est l'analyse des commentaires fournis par les enseignants après la passation des épreuves. Une préoccupation importante est la difficulté d'avoir une épreuve commune aux différents regroupements : les enseignants de A trouvent l'épreuve trop facile (même si aucun de leurs élèves ne l'a réussie complètement), ceux de B la jugent trop difficile pour leurs élèves. En physique par contre, les épreuves portant sur un semestre, l'analyse de la première épreuve permet de modifier certaines questions ou critères de correction en vue de la seconde. Les élèves peuvent regarder la première épreuve, la corriger mais ne la gardent pas car la deuxième épreuve est très semblable à la première. Les concepteurs soulignent que leurs épreuves sont disponibles sur Internet par série d'exercices et que les enseignants peuvent prendre une série et l'utiliser pour leurs évaluations en classe.

Au niveau de ce qu'on pourrait appeler une méta-évaluation (suivi et réflexion sur les épreuves en lien avec la réussite ou l'échec, par exemple, à certaines questions), le plus souvent, le temps et les ressources manquent. La focalisation principale est d'essayer de garder une certaine équivalence d'une année à l'autre, ce qui est plus difficile dans certains domaines que dans d'autres (comme en français au primaire où les domaines et les genres de texte varient d'une année à l'autre).

2.6. Points forts et points faibles des évaluations cantonales ou communes selon les autorités, les directions de l'enseignement et les concepteurs

Nous avons interrogé les différents acteurs pour savoir si l'évaluation commune ou cantonale répondait à tous les besoins de l'institution. Pour plusieurs d'entre eux, l'évaluation répond à de nombreux besoins (trop pour certains) mais pas à tous. Plusieurs relèvent qu'elle ne doit pas être utilisée à des fins de pilotage, même si le plus souvent son utilisation dépasse les fonctions initialement prévues comme l'évolution des acquis des élèves ou la régulation de l'enseignement, notamment lorsqu'on l'utilise à tort ou à raison à des fins de comparaisons entre établissements.

Pour les deux directrices de l'enseignement, il manque actuellement une évaluation-système comme celle qui sera mise en œuvre dans le cadre d'HarmoS et qui permettrait de répondre à la question de savoir si les objectifs sont atteints au niveau institutionnel. L'évaluation cantonale ou commune remplit son rôle défini dans les deux règlements (cf. 2.1) mais par rapport à la régulation du système, le contrôle de l'enseignement et celui du système dépendent d'autres choses que d'une épreuve.

Pour le coordinateur du primaire, ces épreuves sont minimalement standardisées et ne portent que sur les compétences minimales. De plus, elles n'évaluent pas l'ensemble des apprentissages (p. ex. en allemand, l'expression orale n'est pas évaluée). Elles sont censées évaluer prioritairement les acquis des élèves. Elles pourraient participer éventuellement au monitorage ou au pilotage du système mais dans des limites bien définies. D'autres évaluations seraient alors nécessaires. Les épreuves cantonales peuvent fournir des indicateurs de performances des états du système dans la mesure où tous les élèves les passent au même moment. Pour le responsable de la préorientation du CO, les épreuves communes pourraient éventuellement jouer ce rôle si elles subissaient quelques améliorations : de bonnes tables de spécifications, une stabilité d'une année à l'autre permettant une comparabilité, des moyens pour les analyser et donner un retour, etc. Toutefois, il estime qu'il y aurait besoin de plusieurs instruments comme ceux disponibles dans le canton du Valais. L'évaluation commune sert plutôt à certifier. Pour le monitorage, des épreuves de type PISA seraient plus adéquates, notamment pour comparer les établissements.

Pour identifier les points forts et les points faibles du point de vue des acteurs interrogés, nous avons procédé en trois étapes : nous avons d'abord demandé à nos interlocuteurs de nous indiquer les points forts et les points faibles de manière globale, puis d'évaluer de manière générale les évaluations du point de vue de trois types de critères (qualité technique, utilisation, efficience et efficacité) et enfin, pour chacune de ces catégories, nous avons demandé de remplir une grille en détaillant ces critères (cf. annexe 2).

Les points forts

Au niveau des autorités du DIP, on relève la qualité des épreuves cantonales et communes du point de vue de leur conception et de l'équilibre dans ce qu'on recherche. Elles existent depuis plusieurs années et contribuent aux mécanismes de régulation. Elles ont également permis de développer certaines compétences chez les concepteurs.

A *l'école primaire*, il est relevé qu'elles permettent de montrer les attentes institutionnelles et que les résultats permettent une discussion de fond sur les pratiques d'enseignement. Pour le coordinateur, il existe bien un essai d'harmonisation et de cohérence entre les différentes épreuves. La validité interne est satisfaisante. D'autres points positifs sont également mentionnés : l'existence des prétests et l'expérience des formateurs.

Les concepteurs des épreuves relèvent que du point de vue de l'enseignant, il est très appréciable d'avoir une épreuve cantonale permettant de comparer ce qu'il fait dans sa classe, de situer sa classe par rapport à l'ensemble des élèves du canton et donc de se rassurer. Les concepteurs des épreuves de français soulignent la richesse des informations récoltées (p. ex. de la production écrite) et regrettent que l'on n'exploite pas plus certaines données pour définir des attentes de fin de cycle au niveau de l'enseignement. D'autres concepteurs soulignent encore l'importance d'une épreuve qui réunisse tous les élèves et tous les enseignants, ce qui permet de définir des objectifs communs pour tous. C'est également important par rapport à l'orientation au CO. Cela permet aussi une certaine équité pour tous les élèves du canton.

Au *cycle d'orientation*, la directrice de l'enseignement relève plusieurs points positifs : les épreuves communes et cantonales permettent d'attester des savoirs fondamentaux que les élèves doivent avoir acquis à la fin d'une année. Les conditions de passation, les critères de correction, les formes des épreuves font qu'elles sont bien acceptées par les enseignants. Elle sont fiables et crédibles à leurs yeux même si par exemple, la réintroduction d'une évaluation commune ne s'est pas faite sans heurt. Il y a peu d'exemple de résultats « déviants ». Pour le responsable du SEC, la situation varie selon les disciplines : dans certaines, les objectifs sont facilement identifiables, dans d'autres il y a davantage de réticences à ce sujet. Il en va de même par rapport à leur stabilité : par exemple, en allemand, il y a deux épreuves liées à deux méthodes différentes, la future généralisation de la seconde permettra une plus grande homogénéité. Un autre facteur très positif réside dans l'existence même d'EVACOM même si l'application peut encore être améliorée. De manière générale, le responsable de la préorientation relève que l'évaluation commune est une machine qui fonctionne bien et qui produit

beaucoup d'épreuves chaque année. C'est une tradition importante d'avoir une telle évaluation qui permet d'unifier système et pratiques. Par ailleurs, EVACOM est un outil très puissant qui permet une gestion efficace de la saisie des données et des résultats. On peut consulter la base de données par classe, établissement et regroupement (au niveau de la direction générale).

Les commissaires relèvent l'aspect motivant pour les élèves mais aussi pour les concepteurs et les enseignants : les épreuves communes permettent de se situer par rapport à ce qu'on fait en classe. Cela peut également avoir un effet de cohésion sur les trois degrés (par exemple, ce qui est acquis en 7e, ce qui doit être repris en 8e, etc.). Les épreuves communes ont également un effet rassurant sur les enseignants qui savent ce qu'ils doivent faire travailler à leurs élèves au niveau des objectifs et de leur degré d'attente. Le fait que ce soit la même épreuve pour tous est également relevé comme élément positif car c'est l'occasion de tester tous les élèves sur la même épreuve même si les enseignants des différents regroupements peuvent se montrer critiques : Cela s'avère nécessaire pour les élèves si l'on souhaite qu'il y ait une mobilité possible. Selon les commissaires, ce serait une mauvaise idée de baisser le niveau de qualité.

En allemand, les points positifs ont aussi trait au travail en commun pour réaliser les épreuves, le fait d'avoir les mêmes exigences, les effets régulateurs à l'intérieur du groupe. Cela permet aussi d'avoir une crédibilité accrue auprès des parents. Un autre effet positif est d'aider à modifier les pratiques : par exemple, en introduisant l'évaluation de l'expression orale.

Chez les commissaires des épreuves de mathématiques, on relève qu'elles sont respectées : elles intimident les élèves, les maitres les regardent attentivement et cela leur donne des indications sur le niveau d'exigences. Cela leur permet de structurer leur enseignement. De manière générale, elles permettent d'unifier et de fédérer les pratiques. Ils évoquent les épreuves internes réalisées dans la plupart des établissements qui ont des corrélations assez fortes avec les épreuves communes. Par ailleurs, étant donné qu'elles comportent peu de problèmes ouverts, leur correction est assez uniforme.

En physique, on souligne que les épreuves sont considérées comme bien faites. Les enseignants sont contents d'avoir une épreuve tout prête. Ces épreuves donnent également une crédibilité par rapport au niveau d'enseignement postobligatoire : on va vers des attentes qui sont les mêmes pour tous. Les élèves arrivent au collège ou dans les différentes écoles de l'enseignement postobligatoire avec le même *background*.

Les points faibles

Les points faibles relevés par les uns et les autres sont également relativement nombreux. Au niveau du secrétariat général, on estime qu'il faudrait associer des personnes avec des compétences complémentaires (connaissance du contenu disciplinaire, notions de docimologie, etc.) et qu'il faudrait avoir une plus grande cohérence entre les épreuves cantonales de l'école primaire et les épreuves communes du CO au niveau des contenus, des objectifs, des modalités de passation. Une harmonisation serait nécessaire. Il serait souhaitable de fiabiliser ces épreuves notamment au niveau des conditions de passation et de correction. Pour ce dernier point, la correction pourrait être faite en commun au sein d'un établissement, ce qui gagnerait en objectivité. Les analyses et les informations pourraient également être améliorées.

A *l'école primaire*, plusieurs points sont cités par les différents acteurs interrogés :

- des conditions de passation insuffisamment standardisées. Ce point est délicat à régler car pour la SPG, l'évaluation cantonale doit être assurée par les enseignants dans leur classe ;
- l'absence d'enseignants dans les commissions actuellement ;
- les différences d'approches entre les disciplines.

En allemand, les concepteurs relèvent que tout n'est pas mesuré dans l'épreuve cantonale (p. ex. la production orale). D'autres concepteurs évoquent les doubles objectifs de ces épreuves qui leur posent problème. Il y aurait une confusion entre les deux fonctions, celle visant à évaluer les acquis des élèves et celle cherchant à réguler le système.

Au *cycle d'orientation* sont évoquées de manière générale les différences entre disciplines, le problème de l'épreuve commune ou non aux différents regroupements ainsi que l'exploitation des résultats. Les différences entre élèves des différents regroupements sont un problème très difficile à résoudre. Les élèves du regroupement B éprouvent souvent des problèmes de lecture, ce qui a inévitablement des répercussions quand il s'agit d'épreuves papier-crayon, comme c'est le cas des épreuves communes. Pour le responsable du SEC, il est nécessaire pour élaborer l'épreuve de partir d'une table de spécification, de réfléchir à un « squelette » d'épreuve avec des proportions pour les différentes parties ou composantes de l'épreuve. Par ailleurs, il estime que les exercices devraient être porteurs de sens (par exemple dans les épreuves de langue, items portant réellement sur la communication). Il évoque aussi la nécessité de former les commissaires qui le plus souvent arrivent avec de l'expérience (celle de leurs classes) et en acquièrent encore tout au long de l'élaboration des épreuves. Toutefois, au terme de quelques années, les équipes changent. Les prétests ne sont pas encore assez systématiques et le plus souvent servent surtout à tester la formulation des questions. Pour le responsable de la préorientation, la question de la prise en compte des différents regroupements se pose d'une autre manière : il estime important de garder l'idée d'une même épreuve pour tous les élèves avec des barèmes différents afin que les élèves puissent se mesurer à des élèves d'autres classes. Pour l'instant, elles ne sont pas forcément communes : soit l'aide est plus importante pour les élèves de B, soit ce sont des épreuves avec des parties non communes ou des formulations de questions différentes. Il relève des différences entre disciplines : certaines permettent plus facilement une approche « rigoureuse » avec des objectifs clairs, des domaines de compétences spécifiques. De manière générale, il serait nécessaire d'améliorer l'approche en introduisant plus systématiquement des tables de spécification, comme c'est le cas au primaire (cf. annexe 3), qui permettent de savoir plus précisément quels objectifs on teste. Les nouveaux documents de cadrage vont dans ce sens. Par ailleurs, les épreuves devraient être davantage externes pour moduler les résultats de l'enseignant et la correction assumée par d'autres maitres (éventuellement en croisant les classes).

Les commissaires en charge des épreuves de français mentionnent les réactions des enseignants qui se montrent insatisfaits chaque année par rapport au choix du texte et soulignent à nouveau la difficulté de réaliser des épreuves destinées à des élèves des trois regroupements. Il leur faut viser un niveau moyen et être éclectique au niveau des questions pour que tous les élèves soient à l'aise à un moment ou un autre dans l'épreuve en question. Le bachotage provoqué par le stress des épreuves est également un point négatif même si quand il est question de supprimer les épreuves, il y a une levée de bouclier de la part des maitres. Ce bachotage peut être destructeur par rapport à la qualité de l'enseignement notamment pendant les semaines précédant les épreuves. De manière générale, cela peut avoir des effets sur le programme où certains maitres auront tendance à n'enseigner que le champ des épreuves pendant l'année et à laisser de côté certaines parties du plan d'études.

Les commissaires des épreuves de mathématiques mentionnent comme point faible l'exploitation qui est faite des résultats et le manque de ressources pour analyser finement la réussite aux différents exercices. Ils parlent également du revers de la médaille par rapport à la crédibilité de ces épreuves : ce qui n'est pas dans le champ de l'épreuve est moins enseigné que le reste. Par ailleurs, pour eux, il y a des différences entre enseignants concernant le bachotage. Un autre élément est évoqué en lien avec la longueur de l'épreuve (2h) : il s'agit d'une durée inhabituelle pour les élèves de B, incompatible par ailleurs avec leur horaire.

Les commissaires de physique évoquent également les élèves de B. D'après eux, avec ces épreuves il est difficile de valoriser le travail de ces élèves qui sont souvent assez peu scolaires et qui n'aiment pas beaucoup écrire. Ils sont peu à l'aise avec ce type de tâche papier-crayon alors qu'ils réussiraient mieux dans des situations reposant plus sur de l'expérimentation ou de la manipulation. Par ailleurs, les concepteurs des épreuves de physique déplorent le fait qu'ils doivent supprimer des questions intéressantes qui ne seraient pas adéquates pour les élèves de B.

Nos différents interlocuteurs[16] se sont également prononcés au moyen d'une grille sur un certain nombre d'éléments relevant de trois types de critères (cf. annexe 2) :

- les critères de qualité technique,
- les critères d'utilité,
- les critères d'efficience et d'efficacité.

De manière générale, les trois critères sont estimés de manière comparable par les acteurs de l'école primaire, ceux du CO et dans une moindre mesure, les deux membres du secrétariat général. Ces critères, déclinés sous forme d'affirmations, obtiennent en moyenne une valeur se situant entre « partiellement vrai » et « tout à fait vrai ». Seul le troisième critère portant sur l'efficience et l'efficacité rencontre la même adhésion dans les deux niveaux d'enseignement et au secrétariat général.

Concernant la qualité technique, certains éléments donnent lieu à des appréciations plus diversifiées : la validité externe (généralisation des résultats à d'autres contextes), la fidélité ou fiabilité et l'équité. Pour ce dernier critère, plusieurs personnes évoquent les difficultés à ne pas défavoriser les élèves allophones même si les concepteurs s'efforcent d'éviter au maximum les biais.

Concernant l'utilité, plusieurs propositions donnent lieu à des avis contrastés :

- *l'information récoltée est suffisamment complète pour une évaluation qui réponde aux besoins des élèves*. Pour plusieurs personnes interrogées, ce n'est pas le but de l'évaluation cantonale ou commune qui doit être complétée par celle de l'enseignant ;
- *les différents acteurs ou publics considèrent l'évaluation comme valide et non biaisée ;*
- *les personnes chargées de l'analyse des résultats ont les compétences nécessaires pour produire des résultats utiles.* Par ailleurs, cinq personnes n'ont pas donné de réponse. Dans ce cas, cela confirme les réponses apportées concernant les points faibles au sujet de l'analyse des résultats qui, aussi bien au primaire qu'au cycle d'orientation, est peu développée ;
- *les résultats fournis sont faciles à comprendre et informent clairement les différents acteurs sur la manière d'y donner suite ou d'assurer un suivi.* Sept personnes estiment que cette affirmation est peu vraie. Pour les parents, les informations ne sont pas faciles à comprendre et par ailleurs, il n'y a pas d'indications concernant le suivi. On peut sans doute relier cela avec les lacunes et le manque de moyens en matière d'analyse des résultats.

Par contre, pour la diffusion, le « timing », le champ pris en compte par l'évaluation ainsi que la crédibilité, les répondants sont dans l'ensemble assez satisfaits.

Enfin, concernant les critères d'efficacité et d'efficience, deux d'entre eux sont évalués par plusieurs personnes comme peu vraies :

- *l'évaluation est rentable compte tenu du temps et des ressources mises à disposition.* Plusieurs personnes déclarent que le temps consacré à l'élaboration est conséquent. Ils sont souvent partagés : d'un côté, ces épreuves leur paraissent importantes et leur prise en compte n'est pas très élevée et de l'autre, elles ne devraient pas avoir plus de poids étant donné leur caractère ponctuel ;
- *l'organisation actuelle (dégrèvement des enseignants notamment) répond aux objectifs attendus.* Au primaire, plusieurs personnes ont relevé qu'ils regrettaient qu'il n'y ait plus d'enseignants dans les commissions. Au CO, plusieurs personnes ont également parlé du problème des dégrèvements et du fait qu'ils devaient faire deux épreuves en même temps.

[16] Nous avons soumis la grille à 13 personnes ou groupes de personnes. Seuls les deux coordinateurs ou responsables des épreuves cantonales ou communes n'ont pas rempli cette grille, étant donné que nous avons conduit des entretiens exploratoires avec eux.

Les deux tableaux suivants synthétisent le point de vue des autorités, des directions de l'enseignement et des concepteurs concernant les points forts (2.3a) et les points faibles (2.3b) des épreuves et du dispositif en général.

Tableau 2.3a. Points forts des épreuves cantonales et communes selon les autorités, les directions d'enseignement et les concepteurs

	École primaire			CO				DIP
	DGE	Coordinateur EC	Concepteurs EC	DGE	Coordinateur EC	Responsable préorientation	Commissaires EC	SG
Qualité technique								
Évaluation commune: machine qui fonctionne bien et produit beaucoup d'épreuves							●	
Qualité des épreuves (conception et équilibre)			●	●			●	●
Existence de prétests	●	●	●					
Expérience des concepteurs (élaborer des EC contribue à développer également cette expérience)	●	●	●	●				●
Objectifs facilement identifiables (dépend des disciplines) ; présence tables de spécification		●	●		●			
Une certaine stabilité de contenu (différences entre disciplines)					●			
Bonnes conditions de passation, critères de corrections, forme épreuve				●	●			
Bonne validité interne		●			●			
Utilité								
Répond à de nombreux besoins (trop)	●	●	●	●	●	●	●	●
Permet de montrer les attentes institutionnelles ; effet rassurant pour enseignant qui sait ce qu'il doit enseigner	●	●	●				●	
Attester des savoirs fondamentaux que les élèves doivent avoir acquis à la fin d'une année				●				

	École primaire			CO				DIP
	DGE	Coordinateur EC	Concepteurs EC	DGE	Coordinateur EC	Responsable préorientation	Commissaires EC	SG
Utilité								
Réflexion sur les pratiques	●	●	●					
Comparer ce que l'enseignant fait dans sa classe, se situer			●				●	
Contribuer à la régulation de l'enseignement								●
Aider à modifier les pratiques	●	●	●				●	
Permet d'unifier les pratiques						●	●	●
Importance d'une EC qui réunisse tous les élèves et les enseignants → objectifs communs, permet une certaine équité			●		●		●	
Bonnes conditions de passation, critères de corrections, forme épreuve → épreuves bien acceptées				●	●			
Crédibilité par rapport aux autres ordres d'enseignement (p. ex. PO) : on sait ce que les élèves ont fait (attentes les mêmes pour tous) (passage au CO)			●				●	
Effet de cohésion entre degrés (7e-9e)							●	
Aspect motivant pour élèves, enseignants, concepteurs							●	
Utilité par rapport aux parents			●				●	
Efficience/efficacité								
Bonne prédictivité des EC et liens avec les évaluations des enseignants					●		●	
Existence d'EVACOM (application informatique) : outil puissant qui permet une gestion efficace de la saisie des données et des résultats					●	●		

Tableau 2.3b. Points faibles des épreuves cantonales et communes selon les autorités, les directions d'enseignement et les concepteurs

	École primaire			CO				DIP
	DGE	Coordinateur EC	Concepteurs EC	DGE	Coordinateur EC	Responsable préorientation	Commissaires EC	SG
Qualité technique								
Pourraient jouer certains rôles si standardisées (bonnes tables de spécifications, stabilité des EC d'une année à l'autre, moyens pour les analyser et donner un retour)						●	●	
Associer des personnes avec des compétences complémentaires : connaissance du contenu disciplinaire, notion de docimologie, etc.								●
Plus grande cohérence entre primaire et CO par rapport au contenu, objectifs, conditions de passation ; harmonisation								●
N'évaluent pas l'ensemble des apprentissages ; tout n'est pas mesuré		●	● (allemand : prod. orale)					
Contenu et organisation des EC : partir d'un squelette d'épreuve, d'une table de spécification (proportions définies pour les différentes composantes de l'épreuve) ; questions porteuses de sens					●	●		
Nécessité de conditions de passation plus fiables	●	●	●					● (EP)
Nécessité de corrections plus fiables	●	●	●			●	●	●
Différences d'approches et d'EC entre disciplines	●	●	●	●	●			
Problèmes liés à la prise en compte des regroupements A et B (EC commune à tous ou non) : intéresser tous les élèves				●	●	●	●	●
Prétests variables selon les disciplines. Souvent utilisés surtout pour tester formulation des questions					●			
Équipes de commissaires se forment et changent ; formation des commissaires					●	●		

	École primaire			CO				DIP
	DGE	Coordinateur EC	Concepteurs EC	DGE	Coordinateur EC	Responsable préorientation	Commissaires EC	SG
Qualité technique								
Absence d'enseignants dans commissions d'EC	●	●	●					
Utilité								
Répond à de nombreux besoins (trop). Double objectif : confusion (vérifier l'atteinte des objectifs par les élèves et piloter le système)			●					
Manque une évaluation système	●	●		●				
Analyse et exploitation des résultats	●	●	●	●	●	●	●	●
Regret que pas assez exploitées (p. ex. pourraient servir à fixer les attentes de fin de cycle)			●					
Effet de bachotage lié aux EC					●		●	
Provoque du stress chez enseignants et élèves							●	
Réduction du programme au contenu des EC							●	
Efficience/efficacité								
Place que prend l'évaluation (notamment dans carnet)							●	

2.7. Une organisation idéale selon les autorités, les directions de l'enseignement et les concepteurs des épreuves cantonales et communes

Nous avons interrogé les autorités ainsi que les concepteurs des épreuves cantonales et communes sur ce que serait l'organisation idéale pour l'évaluation cantonale et commune. Il est intéressant de constater que pour les uns et les autres, c'est l'organisation existant dans leur ordre d'enseignement qui est la meilleure, moyennant parfois quelques aménagements.

Au niveau du secrétariat général où l'on pilote l'ensemble de la scolarité, on insiste à la fois sur la cohérence à l'intérieur de la scolarité obligatoire et sur une organisation multidisciplinaire. Les personnes s'occupant des épreuves cantonales ou communes devraient posséder des compétences liées à la discipline évaluée mais également des compétences en docimologie (confection des épreuves) et en statistique (analyse des épreuves, constitution de barèmes), en d'autres termes, il faut à la fois des praticiens (importance d'avoir des personnes qui enseignent) et des docimologues. Il serait nécessaire d'avoir la même organisation dans les deux niveaux d'enseignement. Le dispositif devrait être plus

coordonné qu'actuellement au niveau de l'organisation entre acteurs qui s'occupent d'évaluation (primaire, CO et SRED). Actuellement, il existe une commission de liaison EP-CO mais les épreuves sont encore très différentes.

Il faudrait également que cette plus grande coordination s'applique aux différentes phases de l'évaluation : conception, exploitation, diffusion.

Un élément important serait de vérifier (par le SRED) la prédictivité des évaluations cantonales ou communes, et plus particulièrement leur lien avec les résultats de l'année.

La question des regroupements et de la prise en compte des différences de compétences entre élèves de A et de B en particulier est un problème complexe. Les mêmes objectifs pour tous ne signifient pas le même niveau d'acquisition. Pour l'un des membres du secrétariat général, l'épreuve pourrait être commune pour tous les élèves mais l'évaluation finale différente en termes de notes, comme c'est déjà le cas notamment pour le français (qui n'est pas une discipline à niveaux). Il faut trouver le bon niveau de formulation pour éviter que certains élèves se sentent exclus et que d'autres s'ennuient. Pour le second, on pourrait avoir un tronc commun (p. ex. des mathématiques) qui représenterait au moins 50% de l'épreuve. D'après lui, l'organisation idéale devrait également réunir les différents acteurs mentionnés précédemment mais aussi les cadres, les dirigeants, les directeurs d'établissement et les experts de la recherche. Il verrait un dispositif un peu décentré par rapport à la conduite de l'enseignement, de la production du programme et des moyens d'enseignement. Il évoque également la possibilité d'avoir des évaluations cohérentes mais avec des fonctions un peu différentes selon les moments du cursus : une évaluation-bilan à la fin de la scolarité obligatoire ou éventuellement au moment du passage de la 6P au CO. En fin de cycle élémentaire, le type d'évaluation serait différent. On pourrait également imaginer des évaluations différentes selon les degrés : parfois plus par domaines interdisciplinaires, ou parfois par échantillon avec une fonction plus prédictive.

A *l'école primaire,* la majorité des personnes interrogées – didacticiens de la discipline, spécialistes et experts en docimologie – sont relativement satisfaites de l'organisation actuelle, mais regrettent qu'il n'y ait plus d'enseignants dans les commissions pour discuter de la validité des questions, même si parfois les discussions étaient longues et les enseignants avaient quelquefois des exigences plus élevées que les concepteurs.

L'analyse des résultats est également évoquée comme étant actuellement insuffisante. L'introduction d'EVACOM et l'arrivée des directeurs d'établissement peuvent avoir un effet bénéfique sur l'exploitation des résultats.

Pour le coordinateur des épreuves cantonales, les évaluations cantonales gagneraient en crédibilité si elles étaient davantage standardisées, non seulement au niveau de la passation, mais aussi au niveau de la manière de concevoir les questions et de les organiser.

Chez les concepteurs, différents éléments sont relevés selon les disciplines en plus de ceux évoqués précédemment. Ceux du français évoquent la question de la passation (les épreuves pourraient être passées par d'autres personnes). Ils sont favorables au fait que les épreuves soient conçues par des personnes qui connaissent les contenus disciplinaires mais qui n'enseignent pas (question de réflexion et de recul). En allemand, on déplore le manque de cohérence primaire-CO dont on peut supposer qu'il va s'atténuer avec la généralisation d'une même méthode dans les deux ordres d'enseignement. En mathématiques, on souhaiterait une certaine souplesse du cadre. En évaluation, on insiste sur l'intérêt d'avoir une commission composée d'enseignants et de formateurs. Les enseignants d'une part s'enrichissent en participant à l'élaboration des épreuves et d'autre part, apportent leur expérience quotidienne du terrain. Ces enseignants devraient changer régulièrement pour faire évoluer les épreuves. La question de la passation reste un sujet difficile à régler : faire passer l'évaluation par d'autres enseignants ou personnes contribuerait à augmenter le stress des élèves. Pour eux, le mieux serait de standardiser davantage mais d'enlever la note de l'épreuve de la moyenne du trimestre.

Au *cycle d'orientation*, les avis sont un peu différents. Pour la directrice de l'enseignement, ce serait bien d'avoir un service transversal des épreuves cantonales ou communes qui donnerait plus de cohérence au niveau de la fabrication des épreuves mais aussi de la lisibilité au niveau du cursus, ce qui serait cohérent avec le nouveau plan d'études de la scolarité obligatoire, ainsi qu'avec les plans

intercantonaux et nationaux. Selon elle, il faudrait toutefois distinguer opérationnalisation et analyse. L'opérationnalisation reposerait sur des personnes qui connaissent le terrain et les programmes, auxquelles pourraient être associées des personnes externes. Pour l'analyse, il faudrait se donner les moyens de vérifier un certain nombre d'éléments, l'idée étant de faire appel à des ressources extérieures comme le SRED. Il est également intéressant de travailler en collaboration entre les deux ordres d'enseignement comme c'est déjà un peu le cas en français, en allemand et en mathématiques, pour développer la cohérence.

Le responsable du SEC estime que l'institution devrait prendre position par rapport au caractère plus ou moins externe de ces épreuves. Pour lui, même si elles ont été instaurées par les directions de l'enseignement, ce sont des épreuves cantonales et non extérieures, basées sur le même programme et les mêmes manuels. En cela, elles diffèrent des futures épreuves de référence romandes.

Le responsable de la préorientation juge le système actuellement peu adéquat car les moyens investis ne sont pas suffisants. Pour l'instant, l'essentiel des moyens servirait à payer les dégrèvements des enseignants. Pour lui, il serait nécessaire d'avoir des enseignants formés en évaluation et des équipes stables. Un travail avec des personnes spécialistes en construction d'épreuves apportant un éclairage du point de vue méthodologique serait utile. Une collaboration entre acteurs avec des compétences différentes serait une bonne chose.

La principale lacune réside dans le suivi des épreuves et des résultats, ce qui permettrait d'améliorer l'épreuve de l'année suivante. Il souligne également qu'actuellement les compétences en évaluation sont éclatées au CO : d'un côté l'évaluation commune relevant du secteur de l'enseignement, de l'autre la préorientation et la recherche en évaluation qui dépend du service de la scolarité. Pour lui, ce serait bien de réunir ce qui relève de l'évaluation au CO.

Du côté des commissaires dans les différentes disciplines, il ressort que l'organisation actuelle basée notamment sur les dégrèvements d'enseignants leur parait adéquate. En d'autres termes, les épreuves sont faites par des enseignants pour des enseignants. L'important d'avoir une pratique de classe est mise en avant. En allemand, la question de la formation est évoquée, notamment avec l'introduction d'un module consacré à l'évaluation de l'expression orale. En mathématiques, on relève que ce serait bien qu'il y ait quelqu'un qui puisse aller plus loin dans l'exploitation des résultats et qu'il y ait plus de retour avec les autres ordres d'enseignement. En physique, la problématique des prétests est mise en avant : d'une part, leur systématisation et leur analyse, d'autre part, l'intérêt de tester différentes formulations pour construire un savoir à transmettre à l'ensemble des enseignants. Selon eux, l'analyse des résultats devrait être développée.

Tableau 2.4. Organisation idéale du dispositif d'évaluation cantonale et commune selon les autorités et les concepteurs

	École primaire			CO				DIP
	DGE	Coordinateur EC	Concepteurs EC	DGE	Coordinateur EC	Responsable préorientation	Commissaires EC	SG
Qualité technique								
Organisation multidisciplinaire : compétences concernant la discipline évaluée, compétences en docimologie (élaboration de l'épreuve) et en statistique (analyse des épreuves, constitution de barèmes)				●				●
Organisation avec didacticiens formateurs, docimologue + enseignants	●	●	●					
Organisation avec commissaires-enseignants, PG							●	
Prise en compte des différences de niveaux (p. ex. regroupements au CO)								●
Standardiser davantage les EC (conditions de passation, manière de concevoir les questions et les organiser)		●						
Améliorer les conditions de passation			●					
Développer les prétests et leur analyse							●	
Améliorer les conditions de corrections : les faire en équipe au sein d'un établissement pour gagner de l'homogénéité								●
Utilité								
Améliorer l'analyse des résultats	●	●	●			●	●	
Fonctions différentes selon les moments ? (début ou fin de la scolarité)								●
Efficience/efficacité								
Moyens pour la formation des personnes qui élaborent les EC						●	● (allemand)	
Cohérence et coordination à l'intérieur de la scolarité obligatoire			● (allemand)	●				●
Nécessité de vérifier la prédictivité								●

3. Le point de vue des enseignants à propos de l'évaluation cantonale / commune

Le point de vue des enseignants, en tant que principaux utilisateurs des épreuves cantonales/communes, nous est apparu comme indispensable à prendre en compte dans ce rapport. Étant donné les délais impartis, il n'a pas été possible de réaliser, pour l'ensemble du corps enseignant de l'école obligatoire, une enquête par questionnaire qui nécessite à la fois l'élaboration de l'instrument, son envoi, un délai d'attente ainsi que l'entrée et l'analyse des réponses. C'est pourquoi nous avons choisi de procéder par des entretiens. Cette méthode a toutefois des limites car elle ne permet pas d'interroger l'ensemble des enseignants[17] et ne peut concerner qu'un petit échantillon[18]. La sélection a été opérée différemment à l'école primaire et au CO. Au primaire, 6 établissements contrastés sur 91 ont été sélectionnés sur la base de plusieurs critères : composition socioéconomique, proportion d'allophones, taille de l'école et résultats aux EC 2008 dans les trois ou quatre domaines.

Au CO, compte tenu du nombre de disciplines évaluées (6 sur 7 ont fait l'objet d'investigations) et des délais imposés, un choix a dû également être opéré. La direction du CO a choisi de passer par les directions de collèges et le volontariat. Cinq établissements ont répondu à la demande. Il a été décidé de choisir des groupes de deux disciplines dans chaque établissement (sauf un) en croisant disciplines plutôt littéraires (langue d'enseignement et langues 2 ou 3) et plutôt scientifiques (mathématiques, physique, biologie). Neuf entretiens ont pu ainsi être réalisés[19].

Ces entretiens, d'une durée d'environ 45 mn, comportaient deux parties. La première sous forme orale abordait les thèmes suivants : les objectifs de l'évaluation cantonale ou commune, les points forts et faibles de ces évaluations, l'utilisation des épreuves, le lien avec le programme et l'enseignement réalisé durant l'année, les liens avec les résultats des autres évaluations ainsi que des propositions d'amélioration. La seconde partie se présentait sous la forme d'un court questionnaire (cf. annexe 4) composé d'affirmations sur l'évaluation externe (évaluation cantonale ou commune)[20] et d'une question visant à identifier les points forts et les points faibles du point de vue des trois critères (adéquation ou qualité technique, utilité, efficience/efficacité).

Nous allons rendre compte des principaux résultats de l'analyse des entretiens.

[17] En effet, elle est coûteuse en temps car elle comprend la conduite de l'entretien, sa transcription et son analyse.

[18] Étant donné les délais et la manière de procéder différente d'un niveau d'enseignement à l'autre, il n'a pas été possible de constituer un échantillon suffisamment représentatif. Toutefois, la plupart des réponses récoltées confirment celles recueillies dans la première phase auprès des autorités, des directions d'enseignement et des concepteurs des épreuves.

[19] Le premier entretien a donné lieu à un certain nombre de réticences, dues notamment à la demande et au petit effectif de personnes interrogés. Dans ce premier établissement, on a préféré répondre par écrit pour avoir le temps de réfléchir aux différents thèmes.

[20] Cette question a été reprise du questionnaire aux enseignants utilisé dans la recherche EVALEPCOPO (2006) et complétée par des énoncés tirés de l'étude de Ntamakiliro et Tessaro (2002).

3.1. A l'école primaire

Les objectifs

Tableau 3.1. Objectifs des épreuves cantonales d'après les enseignants de l'école primaire

Niveau		Nombre*
Élèves	- évaluer (tous) les élèves, voir leur niveau d'acquisition du programme	4
	- moyen de pression sur les élèves	1
Enseignants	- vérifier si notions abordées	1
	- savoir ce qui est attendu	1
	- se situer, donner des repères	1
Enseignement (régulation)	- harmoniser, unifier	1
	- définir des attentes communes (minimum.) standards	2
	- évaluation externe moins subjective	1
	- photographie uniforme	1
	- effets sur le plan d'études ou les moyens d'enseignement	1
Système	- comparer des écoles, des milieux	2
	- regard de la direction	1

* Le nombre représente la quantité d'établissements ayant mentionné l'élément en question. Chaque unité rend compte de l'avis d'un ou de plusieurs enseignants de l'établissement concerné. Elle sera reprise dans tous les tableaux pour l'école primaire.

On retrouve les deux finalités définies dans le règlement de l'école primaire, à savoir vérifier si les élèves ont acquis le programme et réguler l'enseignement. Une troisième fonction, moins fréquente, s'ajoute : celle du pilotage et du monitorage, inévitablement induite par les résultats de l'ensemble des élèves du canton et ceux par école ou établissement.

Points forts et points faibles

Comme on peut le constater dans le tableau 3.2, si les épreuves sont considérées globalement comme étant de bonne qualité et importantes pour tous, les points faibles sont relativement nombreux, sans doute parce que, de manière générale, il est plus facile de relever des points négatifs. Par ailleurs, comme on pourra l'observer dans les prochaines questions, les appréciations diffèrent selon les disciplines évaluées. Les principaux points forts relevés concernent la qualité ainsi que le caractère commun, voire objectif de cette évaluation. Dans les points faibles, on relèvera un certain consensus concernant le manque de stabilité des épreuves (les enseignants regrettent que pour le français I, d'une année à l'autre, voire d'un degré à l'autre, les composantes ne sont pas les mêmes et en particulier, que la production écrite ne soit pas systématiquement évaluée) ; la manière d'évaluer la production écrite est également mise en avant (beaucoup d'aide est apportée aux élèves qui mobilisent peu leurs capacités de rédaction), les conditions de passation qui reposent sur les enseignants sont également mentionnées comme donnant lieu à des différences entre classes, écoles, etc. La question de l'évaluation en 2P est aussi un point central. Pour certains enseignants, les élèves sont trop jeunes pour être évalués au moyen d'épreuves aussi longues. Le cas de la structuration est pointé (plus dans certaines écoles) : en principe, elle ne compte pas dans la certification. Par ailleurs, dans certaines classes, on n'évalue pas les élèves durant l'année dans ce domaine, étant donné leur niveau de lecture. Dans l'école du milieu socioéconomique le plus favorisé, la structuration ne pose aucun problème. Un autre élément général mis en avant est le moment de l'évaluation qui devrait être plus tard dans

l'année pour que les enseignants aient le temps d'aborder l'ensemble du programme avec leurs élèves (ce qui peut poser un problème de timing au niveau des corrections et du retour des résultats).

Tableau 3.2. Points forts et points faibles des épreuves cantonales d'après les enseignants de l'école primaire

Points forts		Points faibles	
	Nombre		Nombre
Qualité technique			
Bonnes EC	2	Épreuves différentes d'une année à l'autre voire d'un degré à l'autre (différences de composition dans F1	3
Conditions de passation et de correction bien expliquées	2	Manière de tester la production écrite de manière générale (genre textuel plus que capacité à rédiger) et en 2P (bcp trop d'aide)	2
Évaluation moins subjective (que celle des enseignants)	1	Pb/ production écrite absente selon les degrés	1
Contenu : retour sur l'année	1	Niveau pas adapté (EC 2007-2008 F1) (CE texte trop long, CO une seule écoute)	3
Bon reflet de ce qui est attendu	1	Décalage par rapport à la culture des élèves dans certaines écoles (type REP)	1
		Type de barème très large (par rapport aux évaluations de l'année)	1
		Conditions de passation différentes selon classes, écoles	1
		Critères de correction	1
		EC de plus en plus difficiles	1
Utilité			
Important pour élèves et parents (officiel, tous les élèves impliqués)	1	Ce qu'on en fait après (retour sur enseignement)	1
Attentes communes pour tous	1	Trop d'importance pour parents et élèves	1
		Par rapport aux 2P : trop lourd, stressant	2
Divers			
		Trop tôt dans l'année (effet sur le programme couvert par l'enseignant)	1
		Pas de liberté / moment	1
		Caractère normatif de l'EC	1

Utilisation des épreuves

La principale utilisation relevée par la majorité des enseignants est de préparer les élèves à une épreuve plus longue que leurs évaluations habituelles. L'intégration d'exercices dans leurs évaluations régulières peut aller également dans ce sens : elle peut servir à habituer les élèves ou servir à diversifier les exercices utilisés (tableau 3.3).

Une autre fonction mise en évidence se rapporte à la régulation de l'enseignement aussi bien au niveau de ce que savent les élèves que du programme à aborder.

Tableau 3.3. Utilisation des épreuves cantonales d'après les enseignants de l'école primaire

Niveau		Nombre
Élèves	Entrainer les élèves (par rapport à la longueur de l'EC et la formulation des questions) ; préparer, bachoter	6
Enseignants ou enseignement	Intégrer certains exercices dans les évaluations de l'année	3
	Se situer	1
	Voir si on a bien fait le programme	2
	Adapter son enseignement	2
	Régulation de l'enseignement par rapport aux erreurs des élèves	1
	Espèce d'évaluation diagnostique en début d'année (suivante)	1

Propositions d'amélioration

Plusieurs remarques concernent la collaboration entre enseignants et concepteurs au niveau de l'élaboration des épreuves, les conditions de passation, le moment où ont lieu les épreuves, les liens avec le programme, le niveau d'exigences, la stabilité des épreuves du point de vue des composantes (notamment le français I), la présence et l'évaluation de la production écrite, les EC en 2P (et notamment la place de la structuration). Deux thèmes reprennent ceux déjà évoqués par les concepteurs, les DG ou le secrétariat général : conditions de passation et présence d'enseignants dans la phase de conception (notamment pour consultation).

Tableau 3.4. Propositions d'amélioration des épreuves cantonales d'après les enseignants de l'école primaire

Thèmes		Nombre
Remarques générales		
Collaboration	Davantage de collaboration entre enseignants et concepteurs	3
Couverture / programme	EC qui correspondent plus au programme demandé	1
	Meilleure couverture de tout ce qui demandé dans le PE (ex : maths bcp d'espace et peu de technique opératoire dans EC ; pas de production orale)	1
Moment de la passation	Retarder le moment de la passation	2
Conditions de passation et de correction	Corrections, passations : croisement entre classes	2
Barème	Barème, comptage des points, interprétations	1
Niveau d'exigence	Conditions de passation trop sévères (ex pas de relance, une seule écoute, etc.)	2
	Trop à lire en général	2
	Niveau d'exigences (sensibilisation ou plus) (différences de niveaux entre écoles)	1
Type d'épreuve	Type d'épreuve (sélectif) : formatif serait souhaitable	1
Production écrite	Présence de la production écrite dans toutes les EC	1
	Conception de la production écrite (connaître un genre textuel est différent de savoir rédiger)	3
Français I	Même modèle d'une année à l'autre (mêmes composantes évaluées, même genre textuel)	1
Allemand	Élever le niveau de l'épreuve considérée comme trop facile	2
2P	Pb EC 2P : trop difficile en français p. ex. (caractères trop petits, longueur du texte, etc.)	2
	Statut de l'EC de 2P	2
	Pourquoi évaluer le français II en 2P	1

Couverture de l'EC et correspondance avec enseignement de l'année

Le jugement concernant la couverture du programme dans l'EC varie selon les disciplines (pour le français I, certains enseignants évoquent les genres textuels) et les degrés. Les EC sont le plus souvent considérées comme correspondant au programme (*tout à fait* ou *en grande partie*).

Pour ce qui concerne les liens entre l'EC et ce que font les enseignants durant l'année, on observe également des variations selon les disciplines mais également selon les écoles. Une des écoles de milieu très favorisé estime aller au-delà du programme couvert par l'EC pendant l'année. Pour certains, cela correspond bien en maths, pour d'autres, il y a plus de réserve en maths étant donné la trop grande place donnée aux situations-problèmes dans l'épreuve. La correspondance entre EC et ce qui se fait pendant l'année est bien sûr très liée aux pratiques des enseignants.

Estimation de la difficulté (EC 2007-2008)

Pour cette question, nous avons demandé aux enseignants de se référer à l'évaluation de l'année passée (2007-2008).

Tableau 3.5. Estimation de la difficulté des EC 2007-2008 d'après les enseignants de l'école primaire

	2P			4P				6P			
	FI	*FII*	*Maths*	*FI*	*FII*	*Maths*	*All.*	*FI*	*FII*	*Maths*	*All.*
faciles	-	1	-	1	2	1	2	3	1	2	4
moyennes	-	3	5	3	4	5	2	2	5	4	2
difficiles	6	2	1	2	-	-	1	1	-	-	-

N.B. FI = français I, FII = français II, All. = allemand.

Le tableau 3.5, qu'il faut bien sûr considérer avec prudence compte tenu des effectifs, met en évidence un ou deux éléments intéressants : l'unanimité autour de l'EC de 2P en français I considérée comme trop difficile (avec un texte très long en compréhension de l'écrit, et une seule écoute pour la compréhension de l'oral), plusieurs épreuves considérées d'un niveau moyen par la majorité des enseignants des écoles interrogées : maths en 2P et 4P, français II en 6P. Pour les reste, les avis sont un peu plus partagés.

Liens entre résultats EC et évaluations de l'année

Pour faire le lien entre les résultats aux EC et les autres évaluations de l'année, nous avons posé deux questions aux enseignants : les résultats correspondent-ils aux résultats habituels (*pour la majorité des élèves*, *pour une partie des élèves* ou *pas pour les élèves en difficulté*) ? Et comment jugent-ils les résultats aux EC par rapport à ceux de l'année (ils confirment l'évaluation de l'année, les résultats aux EC sont meilleurs, les résultats aux EC sont moins bons) ?

Pour la plupart des domaines, les résultats aux EC correspondent le plus souvent à ceux de l'année pour la majorité des élèves. C'est en allemand que cela correspond le mieux. Deux remarques particulières peuvent être faites : certains enseignants ont insisté sur les résultats « catastrophiques » en français II pour la 2P, notamment parce qu'ils ne l'évaluent pas pendant l'année (en lien avec les connaissances à mettre en place en lecture). Par ailleurs, les enseignants relèvent dans certaines classes l'existence de cas particuliers d'élèves paniquant aux EC ou au contraire réussissant très bien alors que ce n'est pas le cas habituellement.

La seconde question met en évidence des variations entre disciplines. Ainsi, pour le français II, on observe plutôt que les EC confirment les résultats de l'année ; il en va de même pour les maths en 2P. Pour l'allemand, notamment en 6P, l'épreuve qui était considérée comme facile donne lieu à de meilleurs résultats que les évaluations durant l'année. Pour le français, cela parait variable d'une école à l'autre.

Relevons certains éléments intéressants : pour une des écoles (milieu favorisé), quels que soient le domaine et le degré, les résultats obtenus par les élèves aux EC sont meilleurs que ceux de l'année (ce qui peut s'expliquer par les différences d'exigences mises en évidence dans d'autres questions entre le minimum demandé dans les EC et ce qui est exigé et travaillé pendant l'année). Dans une autre école de type REP, on observe l'inverse. Ces décalages peuvent poser problème avec les parents, notamment quand les résultats observés sont meilleurs aux EC.

3.2. Au cycle d'orientation

Les objectifs

Tableau 3.6. Objectifs des épreuves communes d'après les enseignants du CO

Niveau		Nombre*
Élèves	- évaluer, tester si seuils minima acquis sur l'ensemble du canton	1
	- s'assurer que tous les élèves ont acquis les points essentiels ou vérifier l'atteinte / PE	4
	- bilan des connaissances acquises	1
	- habituer les élèves à une situation d'évaluation-bilan de type examen	1
	- point de comparaison entre élèves	1
	- sélection et orientation élèves par niveau, par rapport à l'entrée en apprentissage	3
	- motiver les élèves face à ce contrôle officiel (travail, voir ses lacunes, se situer)	2
Enseignants	- situer ses élèves par rapport à la moyenne cantonale	1
Enseignement (régulation)	- harmoniser, unifier l'enseignement	4
	- baliser, harmoniser les degrés d'exigence	3
	- régulation de l'enseignement	2
	- noter objectivement les élèves (tests et critères communs)	1
	- harmoniser les progressions (degrés)	1
	- rendez-vous, point de repère annuel	1
	- avoir un minimum de connaissances communes pour l'ensemble des élèves du canton	1
Système	- fournir un point de vue global des acquis des élèves au niveau de l'école, du canton	1
	- avoir une base de comparaison entre CO par niveau	1
	- comparer les écoles entre elles	1
	- vérifier pour la direction que tous les enseignants abordent les mêmes thèmes	1

*Le nombre se réfère à un groupe de disciplines d'un établissement (il peut émaner d'un ou de plusieurs enseignants de ce groupe). Dans tous les tableaux qui concernent l'opinion des enseignants du CO, la même unité sera reprise.

Dans les grandes lignes, comme au primaire, on retrouve les deux principales finalités énoncées dans les règlements : évaluer les acquis des élèves et réguler l'enseignement. Comme au primaire, la dimension « pilotage du système » est également évoquée mais dans une moindre mesure.

Certains éléments tels que la sélection, l'orientation ou l'accent sur la vérification que tous les élèves ont atteint les objectifs sont mentionnés alors qu'ils ne l'étaient pas ou peu sous cette forme au primaire.

Points forts et points faibles

Les points forts

Tableau 3.7a. Les points forts des évaluations communes d'après les enseignants du CO

		Discipline	Nombre
Qualité technique			
Contenu / programme	EC testent les trois compétences	allemand	1
	Fidèles aux objectifs annoncés, programme complet	physique	1
Organisation du contenu	EC bien faites, bien équilibrées (+ contenu et niveau)	anglais, mathématiques, physique	3
	EC bien faites : présence des trois composantes (C, ML et L). Donnent des infos, bon équilibre entre les trois domaines	français	1
	Unité thématique	allemand	1
Clarté épreuves et critères correction	Clarté des épreuves et des critères de correction	anglais	1
Utilité			
Effets des conditions de passation ou de l'EC sur élèves	Conditions de passation : habituent les élèves à avoir des épreuves longues, à fonctionner de manière autonome (sans questions aux enseignants)	français	1
	Habituent / préparent les élèves à être évalués sur un champ large	allemand	1
	EC motivent les élèves pour les révisions (difficile quand pas d'EC comme en 7e)	anglais	1
Utilité par rapport à l'harmonisation	Permet de mettre des lignes directrices dans PE flou	biologie	1
	EC créent une unité , une cohésion au niveau de l'enseignement	français	1
	Important qu'il y ait une EC, tout cohérent avec PE et ME	mathématiques	1
Connaître les acquis	Retour p. ex. : permet de savoir ce que savent les élèves	mathématiques	1
Effets sur les évaluations internes	Varier l'évaluation, manière d'aborder les questions	mathématiques	1
Divers	Énorme mérite : exister	français	1

Les points faibles

Tableau 3.7b. Les points faibles des évaluations communes d'après les enseignants du CO

		Discipline	Nombre
	Qualité technique		
Prétests	Absence de prétests, faute de moyens	allemand	1
Consignes et/ou correctifs	Pb récurrent : consignes peu claires	allemand	1
	Viennent d'être réintroduites (encore bcp d'erreurs au niveau du vocabulaire des énoncés et des correctifs)	biologie	1
Barèmes	Pb de barème et d'attribution de points ; difficile de mettre les points	allemand, biologie, mathématiques	3
	Pb au niveau des barèmes basés sur la réussite des élèves (50% des élèves pour déterminer le seuil du 3.5 et non sur le nombre de points) différent des éval. de l'année	mathématiques	1
	Barèmes pas adaptés à certains niveaux (8R et 9B)	mathématiques	1
Adaptation aux différents niveaux ou regroupements	EC pas adaptées aux B car trop scolaires. Démotivent et stressent les élèves	allemand	1
	Pas adaptées au niveau des élèves : trop faciles pour les A (trop courte en temps : biol.), trop difficiles pour les B	français, biologie	2
Stabilité de l'épreuve	Manque d'innovation (trop similaire d'une année à l'autre)	anglais	1
Contenu de l'épreuve / programme	Compréhension écrite difficile	anglais	1
	Différences de centration PE et EC : PE centré sur compétences et EC sur connaissances	français	1
	Contenu : pas forcément les mêmes points de centration que dans le PE, poids des différents domaines (notions, sous-domaines, etc.)	mathématiques	1
	Savoir-faire prime sur contenu	biologie	1
	Production écrite : point faible dans l'EC ou EC production écrite ne teste pas vraiment les compétences (trop élémentaire, très directive)	français	2
Type d'évaluation	QCM trop réducteurs	français	1
	Pas représentative du niveau des élèves (partie avec QCM et partie de réflexion), composition de l'épreuve (démarches / contenus)	biologie	1
	Ne reflète pas bien la manière d'enseigner (communication difficile à tester par des ex. pointus), ne reflètent pas les réelles capacités des élèves	anglais	1
Correction	Correction par les enseignants (proposition de croiser les classes)	physique	1

Utilité			
Retour des résultats	Résultats globaux par compétences moins pertinents que par ex pour localiser les difficultés des élèves	anglais	1
Liens avec les évaluations des enseignants	Enseignants ne se reconnaissent pas dans ce type d'évaluation	français	1
Apport	Quel est l'apport ? Trop général / grammaire, ne signifie rien	français	1
Liens avec plan d'études	EC trop calquées sur PE (pas d'accord avec la conception du Plan d'études)	biologie	1
Efficacité/efficience			
Prise en compte de l'épreuve	Trop de poids attribué aux EC	français	1
	Épreuve compte comme les autres mais angoisse des élèves	physique	1
	Validation du cours option pas le même poids au collège	physique	1
Autres			
Surcharge	Surcharge de travail (correction et saisie résultats)	anglais	1

Comme pour la plupart des questions, les réponses sont forcément liées à la discipline concernée. On peut toutefois trouver des remarques transversales aux disciplines : les épreuves ont le mérite d'exister ou elles sont bien faites, bien conçues (relevé dans plusieurs disciplines : anglais, maths, physique, français). Les autres points positifs relèvent plus d'une discipline particulière. Les points faibles sont assez nombreux. Deux semblent se dégager : la question des barèmes (dans plusieurs disciplines, ils sont calculés a posteriori sur la distribution des résultats des élèves, le 3.5 étant fixé en fonction des résultats de 50% des élèves) ; l'autre point relevé très fréquemment, que les disciplines soient à niveaux ou non, c'est la prise en compte des regroupements A et B (l'épreuve étant trop facile pour les uns, trop difficile ou inadaptée pour les autres). Souvent, pendant les cours, les enseignants procèdent différemment avec les élèves de B. La prise en compte du programme est perçue différemment selon les disciplines : dans certains cas, on estime que l'esprit (compétences *vs* connaissances) est différent (p. ex. du français), ou que le poids des différents domaines n'est pas le même, ou encore que cela pose problème car très lié au plan d'études (vécu comme flou ou problématique).

Utilisation des épreuves

Tableau 3.8. Utilisation des évaluations communes d'après les enseignants du CO

Niveau		Nombre
Élèves	Entrainer les élèves (par rapport à la longueur de l'EC et la formulation des questions) (utilisation pour la révision) ; préparer, bachoter	4
	Utilisation d'exercices (entrainement) éventuellement à la maison	3
	Déstresser les élèves, les habituer aux consignes des épreuves la gestion du temps	2
Enseignants ou enseignement	EC en libre accès sur internet (y compris pour élèves)	1
	Référence	1
	Donner des idées pour situer leurs élèves par rapport à la moyenne cantonale	1
	Intégrer certains exercices dans les évaluations de l'année ou s'en inspirer	2
	Adapter son enseignement (pour préparer les élèves aux épreuves)	1
	Régulation de l'enseignement (fil directeur, modèle, faire comprendre aux élèves ce qu'ils doivent acquérir)	1
	Peu ou pas d'utilisation (pour raisons diverses)	3

Comme au primaire, les deux principales utilisations ont trait à l'entrainement des élèves du point de vue de la longueur des épreuves et de la formulation des questions (EC complète ou partielle) ou à la régulation de leur enseignement d'une manière ou d'une autre (en intégrant des exercices dans leur évaluation, en adaptant leurs pratiques d'enseignement). Certains enseignants ont déclaré ne pas les utiliser soit par manque de temps, soit parce qu'elles ne correspondent pas à leurs pratiques (ou qu'ils n'en avaient pas le droit comme en biologie car il s'agissait de la deuxième année).

Propositions d'amélioration

Tableau 3.9. Propositions d'amélioration des évaluations communes d'après les enseignants du CO

Thèmes		Discipline	Nombre
Remarques générales			
Prétests	Épreuves devraient être prétestées	allemand	1
Corrections	Corrections faites par d'autres enseignants (croisement de classe)	allemand, physique	2
Prise en compte des différences de niveaux	Épreuves différentes A et B : ne pas tester les B de la même manière ni sur tous les domaines (p. ex. compréhension seulement en allemand)	allemand, français, mathématiques	5
	Épreuve par regroupement en fonction du niveau des élèves (p. ex. plus de rédaction, de questions d'interprétation, etc. pour A)		
	EC inadaptées aux élèves les plus faibles (C) (très différent du travail fait en classe avec eux) ; épreuve pour 9e atelier : non-sens)		
	Nécessité d'un barème adapté aux élèves de 8R et 9B		
Prise en compte de l'EC	Mise en évidence dans carnet pour les B : problème car meilleurs résultats EC (décalage/résultats en classe où l'on prend aussi en compte l'investissement des élèves et une part plus grande donnée à la pratique	physique	1
	Prise en compte de l'EC : même poids qu'une autre épreuve et pas de mise en évidence dans le carnet de l'élève	mathématiques	1
Barèmes	Fixer les barèmes à l'avance en fonction des objectifs ; barème a priori en fonction de ce qu'on attend concernant les objectifs (et non seuil de réussite fixé sur la réussite de 50% des élèves)	français mathématiques	1
	Constitution et diffusion du barème plus tôt	mathématiques	1
Moment de la passation	Moment de l'année :	mathématiques	1
	- soit champ plus restreint ou épreuve plus tardive (pour pouvoir finir le programme)		
	- intervenir plus tôt avec un champ restreint		
Types de compétences évaluées	Évaluer des compétences plus générales (utilisation d'une langue) et pas seulement points précis (évaluer une compréhension écrite qui cible moins des « niches » de connaissance	anglais	2
Répartition des différents champs dans l'EC	Réfléchir à l'adéquation des % des différents *skills* (leur répartition dans l'EC)	anglais	1
	Meilleure prise en compte des % du PE (des différents domaines et des compétences)	mathématiques	2
	Déséquilibre entre numérique et géométrique (8e-9e) au détriment de la géométrie		
Gestion des élèves	Gestion des élèves absents : la même dans tous les CO (séances de rattrapage ou non)	mathématiques	1
Utilisation	Préciser l'utilisation qu'en fait la direction	allemand	1

Contenu des EC / discipline			
Contenu / types de connaissances	Mieux définir les objectifs du programme à tester	mathématiques	1
	Cibler sur les compétences (en lien avec le PE) et pas sur le notionnel et le bachotage	français	1
	Incohérence entre EC et PE basé sur compétences	français	1
	Revoir répartition domaines (p. ex. trop basé sur anatomie et pas assez sur physiologie)	biologie	1
	Mettre plus de production écrite, faire des exercices rédactionnels	français	1
	Connaissances plus synthétiques	biologie	1
	Éviter de tester des domaines de façon lacunaire (p. ex. discours indirect abordé avec deux phrases)	français	1
	Mettre plusieurs types de textes dans la même épreuve	français	1
	Tester les compétences de base, se baser sur connaissances mobilisables et pas en voie d'acquisition	mathématiques	1
Niveau d'exigences	Augmenter le niveau de complexité pour montrer les différences entre élèves	biologie	1
Type de question pour A	Épreuve pour A : supprimer les QCM (remplacer par courte réponse)	biologie	1
Stabilité des EC	Structure reproductible d'année en année (exercices-types)	mathématiques	1
Correction	Revoir système de correction trop rigide	français	2
	Pb d'évaluation (compréhension générale des élèves)	mathématiques	
Autres	Regret que pas d'EC en 7e car mobilise les élèves jusqu'à la fin de l'année	anglais	1
	Augmenter la relecture des EC	mathématiques	1

Un certain nombre de propositions d'amélioration peuvent être considérées comme générales dans le sens qu'elles ne portent pas sur le contenu strict de l'épreuve : vient en tête la gestion des différents niveaux. Dans la plupart des disciplines, qu'elles soient à niveaux comme les mathématiques ou l'allemand, ou les mêmes pour tous comme le français et l'anglais, la majorité des enseignants évoquent la difficulté de tester les mêmes choses chez tous les élèves. Dans certains cas, il est précisé qu'avec les élèves en difficulté (de B voire de C), le travail effectué en classe est très différent pour les remotiver, notamment, ou tenter de combler les lacunes. Deux autres *leitmotive* reviennent relativement fréquemment : les conditions de correction (croiser les classes est une proposition) et la manière de constituer les barèmes a posteriori sur le 50% de réussite des élèves et non a priori en fonction des objectifs (cette manière existe notamment en physique). La prise en compte de l'EC dans le carnet est également évoquée. Pour plusieurs enseignants, elle devrait avoir le même poids que les autres épreuves et non figurer à part dans le carnet, notamment parce qu'elle peut être différente de la moyenne annuelle.

Le moment auquel se déroulent les épreuves est aussi sujet à discussion : se situer plus tôt dans l'année avec un champ restreint ou au contraire avoir lieu plus tard, pour que le programme soit bouclé.

La nécessité d'épreuves prétestées est également mentionnée, mais assez rarement.

Les remarques plus liées au contenu sont nombreuses, portant sur la représentativité des différents domaines dans l'épreuve (mathématiques, anglais, français) aussi bien que sur la manière de le faire (trop notionnel et pas assez sur les compétences).

La question du niveau de complexité ou d'exigence de l'épreuve donne lieu à des propositions contrastées : l'augmenter pour mettre en évidence des différences entre élèves ou ne tester que des compétences mobilisables et non en cours d'acquisition.

Couverture de l'EC et correspondance avec enseignement de l'année

Ces deux questions donnent lieu à des réponses très différenciées selon les personnes, à l'intérieur d'une même discipline. Pour certains, les EC couvrent bien le programme comme en allemand ou en physique, pour d'autres très partiellement. Il en va de même pour ce qui concerne les liens entre l'EC et ce qu'ils font en classe durant l'année. On observe un certain consensus en français où les enseignants sont d'avis que l'EC ne correspond que partiellement, voire pas du tout à leur enseignement ou à leur niveau d'exigences. Cela met en évidence la difficulté de toute épreuve à tester des compétences complexes.

Estimation de la difficulté (EC 2007-2008)

Bien qu'il soit difficile de mettre en évidence des tendances étant donné le faible effectif de répondants et le nombre de disciplines prises en compte, on peut observer une certaine constance dans les réponses : pour la plupart, les EC sont considérées comme faciles (notamment en français, allemand, anglais voire biologie) pour les élèves de A ou moyennes (mathématiques et physique) et difficiles pour les B (ensemble des 6 disciplines). Les enseignants font là encore des commentaires concernant la nécessité d'avoir des épreuves différentes pour les deux types de regroupement (tendance encore accentuée avec les élèves de C).

Liens entre résultats EC et évaluations de l'année

S'il est difficile de mettre en évidence des tendances, on relèvera que le plus souvent les résultats aux EC correspondent à ceux de l'année pour une bonne partie des élèves de A voire de H et pour une partie des élèves de B. Quand on pose la question de savoir si les résultats aux EC confirment, sont meilleurs ou sont moins bons que les résultats aux évaluations de l'année, les réponses sont très contrastées : en français, mathématiques et physique, la tendance est qu'ils confirment ceux de l'année pour les élèves de A et qu'ils sont meilleurs en mathématiques et en français pour les élèves de B. En physique, ils sont moins bons en B. En anglais et en allemand, ils ont plutôt tendance à être meilleurs pour les élèves de A et moins bons pour les élèves de B. En biologie, ils sont meilleurs aussi bien pour les élèves de A que de B.

3.3. Opinions générales sur l'évaluation externe (commune ou cantonale), identification des points forts et faibles

Nous avons également demandé aux enseignants de remplir une grille comportant un certain nombre d'affirmations concernant l'évaluation externe sur lesquels ils devaient se prononcer (cf. annexe 4).

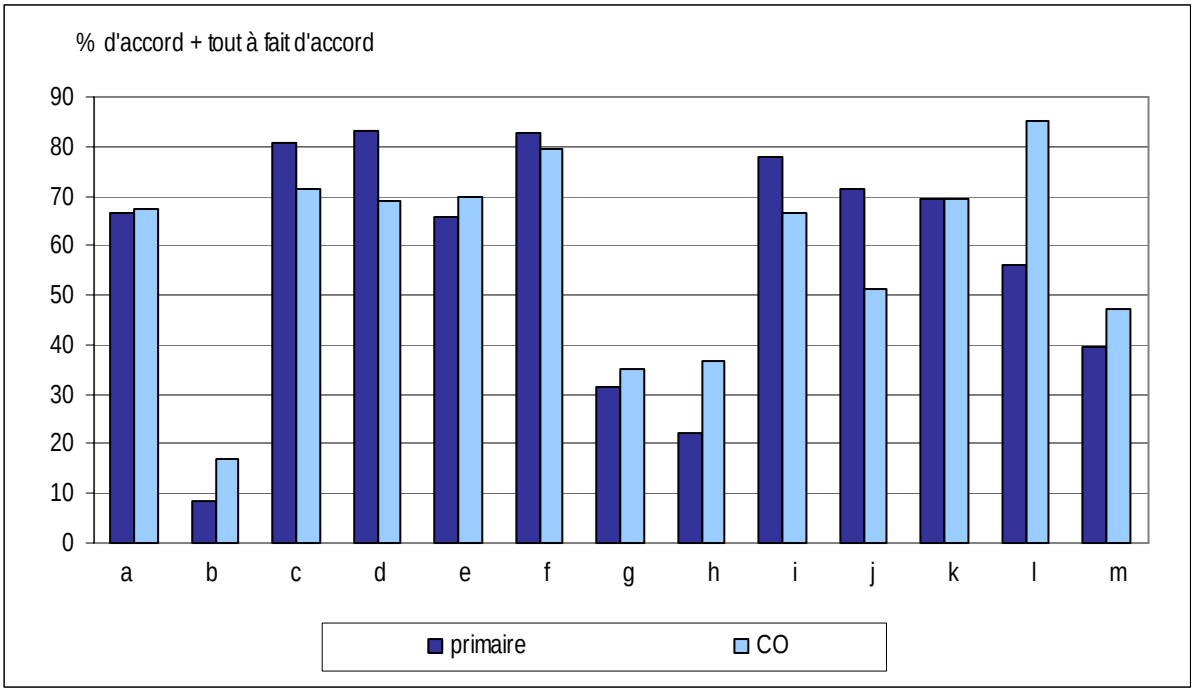
De manière générale, on observe (graphique 3.1) assez peu de différences entre les deux ordres d'enseignement, sauf pour trois items : *l'évaluation externe entraine un certain bachotage* (1l), *l'évaluation externe oblige (ou a pour effet) de limiter le champ de l'enseignement au contenu des épreuves (ou de l'évaluation)* (1m) et *l'évaluation externe ne tient pas compte de ce qui est vraiment enseigné en classe* (1h). Pour les deux premières affirmations, les enseignants du CO ayant répondu font preuve d'un accord sensiblement plus élevé[21]. Pour la troisième, si les enseignants ont plutôt tendance à ne pas être d'accord avec cette affirmation concernant le rapport entre ce qui est enseigné en classe et ce qui est évalué dans l'évaluation externe, ceux du CO sont près de 40% à être plutôt d'accord, voire tout à fait d'accord.

Par ailleurs, pour certains éléments, on observe des différences d'opinions à l'intérieur d'un ordre d'enseignement. Ainsi, à l'école primaire, chez les enseignants interrogés, il ne semble y avoir de

[21] Si l'on compare les moyennes pour chaque item, ce sont les seules affirmations pour lesquelles les différences sont statistiquement significatives.

consensus sur les éléments suivants : *l'évaluation externe sert à évaluer les enseignants* (1g), *l'évaluation externe oblige ou (a pour effet) de limiter le champ de l'enseignement au contenu des épreuves (ou de l'évaluation)* (1m). On observe le même type de phénomène dans l'échantillon d'enseignants du CO concernant les affirmations suivantes : *l'évaluation externe contribue à harmoniser les pratiques d'évaluation entre enseignants* (1c), *l'évaluation externe permet une certaine objectivité parce qu'elle tient compte de ce qui est attendu pour un degré ou un cycle donné* (1d), *l'évaluation externe ne tient pas compte de ce qui est vraiment enseigné en classe* (1h), *l'évaluation externe oblige ou (a pour effet) de limiter le champ de l'enseignement au contenu des épreuves (ou de l'évaluation)* (1m). Ces différences peuvent être entre autres attribuables à des différences de perception liées à la discipline enseignée et à l'EC s'y rapportant.

Graphique 3.1. Opinion sur l'évaluation externe



Légende du graphique

a) L'évaluation externe peut être génératrice d'idées nouvelles.

b) L'évaluation externe n'est pas pertinente car seul l'enseignant est à même d'évaluer les apprentissages de ses élèves parce qu'il les connaît bien.

c) L'évaluation externe contribue à harmoniser les pratiques d'évaluation entre enseignants.

d) L'évaluation externe permet une certaine objectivité parce qu'elle tient compte de ce qui est attendu pour un degré ou un cycle donné.

e) L'évaluation externe permet d'évaluer et de réguler le système.

f) Il est important d'avoir, à côté de l'évaluation régulière réalisée par l'enseignant, une évaluation commune à tous les élèves d'un degré ou pour un cycle car elle donne des repères précis et objectifs.

g) L'évaluation externe sert à évaluer les enseignants.

h) L'évaluation externe ne tient pas compte de ce qui est vraiment enseigné en classe.

i) L'évaluation externe est utile pour s'assurer que les objectifs fondamentaux sont atteints.

j) L'évaluation externe permet de savoir précisément ce qui est attendu.

k) L'évaluation externe a pour effet de stresser les élèves.

l) L'évaluation externe entraine un certain bachotage.

m) L'évaluation externe oblige (ou a pour effet) de limiter le champ de l'enseignement au contenu des épreuves (ou de l'évaluation).

Voici les affirmations pour lesquelles les enseignants font preuve d'un accord dépassant les 60% (et allant même jusqu'à 80%) :

1c) *L'évaluation externe contribue à harmoniser les pratiques d'évaluation entre enseignants*

1d) *L'évaluation externe permet une certaine objectivité parce qu'elle tient compte de ce qui est attendu pour une degré ou un cycle donné*

1f) *Il est important d'avoir, à côté de l'évaluation régulière réalisée par l'enseignant, une évaluation commune à tous les élèves d'un degré ou pour un cycle car elle donne des repères précis et objectifs*

1i) *L'évaluation est utile pour s'assurer que les objectifs fondamentaux sont atteints*

et dans une moindre mesure :

1a) *L'évaluation externe est génératrice d'idées nouvelles*

1e) *L'évaluation externe permet d'évaluer et de réguler le système*

1j) *L'évaluation externe permet de savoir précisément ce qui est attendu*

1k) *L'évaluation externe a pour effet de stresser les élèves.*

La plupart des éléments positifs apportés par l'évaluation externe rencontrent un certain accord aux deux niveaux d'enseignement. Il est intéressant de souligner que parmi les propositions sur lesquelles les enseignants semblent le plus d'accord, figure un élément négatif : le stress occasionné par l'évaluation externe auprès des élèves. 70% des enseignants du primaire et du CO ayant répondu sont d'accord, voire tout à fait d'accord avec cette affirmation.

Par ailleurs, si l'on regroupe toutes les affirmations positives de la question, on observe un degré d'accord moyen quasi équivalent chez les enseignants des deux ordres d'enseignement, atteignant 1.8 (sur une échelle de 0 à 3).

Par contre, pour les éléments négatifs, le rejet des affirmations est plus élevé chez les enseignants primaires que chez ceux du CO (1.2 pour le primaire, 1.45 pour ceux du CO). En d'autres termes, la moyenne se situant à 1.5, les enseignants du CO seraient « presque » d'accord avec un certain nombre d'affirmations négatives se rapportant à l'évaluation externe. On retrouve certains résultats déjà observés dans une étude précédente où tous les enseignants étaient interrogés. (EVALEPCOPO, 2006).

De manière générale, quatre affirmations négatives par rapport à l'évaluation externe font l'objet d'un rejet important de la part des enseignants des deux ordres d'enseignement :

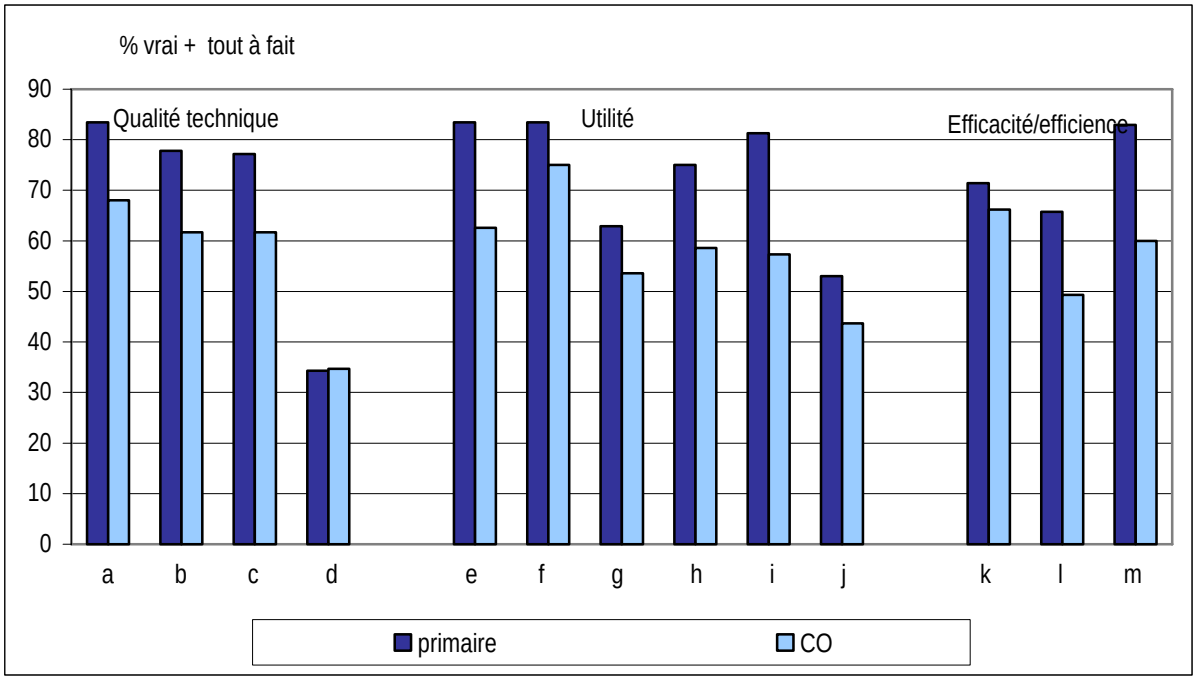
1b) *L'évaluation externe n'est pas pertinente car seul l'enseignant est à même d'évaluer les apprentissages de ses élèves parce qu'il les connaît bien*

1g) *L'évaluation externe sert à évaluer les enseignants*

1h) *L'évaluation externe ne tient pas compte de ce qui est vraiment enseigné en classe*

1m) *L'évaluation externe oblige (ou a pour effet) de limiter le champ de l'enseignement au contenu des épreuves ou de l'évaluation.*

Graphique 3.2. Identification des points forts et faibles de l'évaluation externe



Légende du graphique

Adéquation / qualité technique

a) L'évaluation cantonale ou commune constitue une mesure adéquate des acquis des élèves.

b) Les informations récoltées sont précises et cohérentes.

c) L'évaluation cantonale ou commune est objective : d'autres évaluateurs arriveraient aux mêmes conclusions.

d) L'évaluation cantonale ou commune est favorable ou défavorable de la même manière à tous les élèves (garçons/filles, francophones/allophones, élèves de culture et de milieux socioéconomiques différents).

Utilité

e) Les résultats sont pertinents pour les destinataires de l'évaluation (enseignants, parents, élèves, ….).

f) L'évaluation cantonale ou commune a pris en compte ce qu'il y a d'important et de significatif au niveau du programme.

g) L'information récoltée est suffisamment complète pour une évaluation qui réponde aux besoins des élèves.

h) L'évaluation cantonale ou commune est valide et non biaisée.

i) Les résultats sont fournis aux acteurs concernés quand ils en ont besoin.

j) Les résultats fournis sont faciles à comprendre et informent clairement les différents acteurs sur la manière d'y donner suite ou d'assurer un suivi.

Efficacité/efficience

k) Compte tenu du temps et des ressources utilisés, la prise en compte des EC dans l'évaluation des élèves est adéquate.

l) Les épreuves communes/cantonales sont un bon prédicteur de la réussite des élèves pour une discipline donnée.

m) Les résultats des élèves aux épreuves communes/cantonales sont cohérents avec ceux des autres évaluations de l'année effectuées par les enseignants.

Dans cette deuxième partie du questionnaire, nous avons repris les catégories de la grille soumise aux autorités et aux concepteurs des épreuves (cf. chapitre 2).

La grande majorité des propositions sont considérées comme partiellement vraies, voire tout à fait vraies par les enseignants des deux ordres d'enseignement, à l'exception de deux items : *l'évaluation cantonale ou commune est favorable ou défavorable de la même manière à tous les élèves (garçons/filles, francophones/allophones, élèves de culture et de milieux socioéconomiques différents)* (2d) et dans une moindre mesure *les résultats fournis sont faciles à comprendre et informent clairement les différents acteurs sur la manière d'y donner suite ou d'assurer un suivi* (2j). Dans l'ensemble, on observe assez peu de différences entre les réponses des enseignants primaires et ceux du CO sauf pour deux items : *les résultats sont fournis aux acteurs concernés quand ils en ont besoin* (2i) et *les résultats des élèves aux épreuves communes/cantonales sont cohérents avec ceux des autres évaluations de l'année effectuées par les enseignants* (2m). Dans les deux cas, les enseignants primaires sont plus de 80% à trouver ces affirmations tout à fait, voire partiellement vraies alors que ce n'est le cas que d'environ 60% des enseignants du CO.

Par ailleurs, pour de très rares items, on observe des différences à l'intérieur d'un niveau d'enseignement : par exemple sur l'équité des EC, les enseignants ne semblent pas d'accord entre eux, aussi bien à l'école primaire qu'au CO.

Concernant les trois critères (adéquation/qualité technique, utilité et efficience/efficacité), on observe peu de différences entre les deux ordres d'enseignement, comme on peut le constater dans le tableau 3.10. Dans l'ensemble, pour les trois critères, les enseignants ont un avis se situant autour de la moyenne : au niveau de l'adéquation, les affirmations sont considérées en moyenne comme partiellement vraies, l'utilité et l'efficacité/efficience étant légèrement supérieures.

Tableau 3.10. Degré d'accord moyen des enseignants des deux ordres d'enseignement concernant les points forts et faibles de l'évaluation cantonale/commune

Ordre d'enseignement		Adéquation	Utilité	Efficacité/efficience
CO	Moyenne	1.54	1.61	1.62
	N	76	76	75
	Écart-type	0.54	0.62	0.62
primaire	Moyenne	1.64	1.83	1.79
	N	36	36	36
	Écart-type	0.55	0.42	0.54
Total	Moyenne	1.57	1.68	1.67
	N	112	112	111
	Écart-type	0.55	0.57	0.60

Synthèse

Pour ce qui concerne les objectifs de l'évaluation cantonale ou commune, les enseignants du primaire comme du CO attribuent aux épreuves les deux principaux objectifs annoncés par l'institution, à savoir vérifier si les élèves ont acquis les objectifs attendus et réguler l'enseignement. A cela, ils ajoutent parfois une dimension de pilotage et de monitorage, sans doute induite par les indications de moyennes cantonales. Au CO, la dimension d'orientation est mise en évidence, ce qui n'est pas le cas au primaire, même en 6P, moment du passage au CO.

L'identification des points forts et des points faibles est bien sûr liée non seulement au degré mais également à la discipline concernée. Dans les points positifs, les enseignants de l'école primaire relèvent une bonne qualité des épreuves, qui reflètent bien ce qui est attendu et qui permet d'avoir des attentes communes pour tous. Pour certains, elles sont vécues comme positives parce qu'elles sont moins subjectives que celles réalisées par l'enseignant dans sa classe. Les conditions de passation et de correction sont jugées comme bien expliquées. Cela ne les empêche pas d'être également assez critiques sur un certain nombre de points : le manque de stabilité des épreuves[22] (au niveau des domaines testés par exemple), le fait que certains domaines ne sont pas évalués, les conditions de passation, la question de l'évaluation en 2P (et notamment en structuration), le moment de l'évaluation considéré comme étant trop tôt dans l'année ce qui a des effets sur le programme couvert en classe, le niveau d'exigences (les perceptions sont variables selon les écoles et la composition socioéconomique).

Au CO, les avis sont encore davantage liés aux disciplines, les enseignants étant interrogés par rapport à l'épreuve s'y rapportant. Certaines remarques ont toutefois une portée générale. Parmi les points positifs, les enseignants relèvent qu'elles permettent d'habituer ou de préparer les élèves à être évalués sur un champ plus large et de fonctionner de manière autonome. Elles sont bien faites, équilibrées du point de vue de leur contenu (anglais, mathématiques, physique, voire français). Elles créent une certaine cohésion au niveau de l'enseignement. Dans plusieurs disciplines, elles sont considérées comme fidèles au programme. Les points faibles relevés sont très nombreux et souvent très dépendants de la discipline considérée. Certains sont toutefois repris par des enseignants de différentes disciplines : le problème des barèmes a posteriori davantage basés sur les résultats des élèves et ne tenant pas forcément compte de ce qu'on devrait attendre des élèves du point de vue des objectifs du plan d'études (en physique, la situation est différente), la difficulté extrême d'avoir des épreuves communes aux différents niveaux (pour la majorité, elles ne sont pas adaptées aux élèves de B car trop scolaires et différentes de ce que les enseignants font en classe, et trop faciles pour les A). Un certain nombre de critiques concernent le contenu de l'épreuve (connaissances pointues évaluées dans l'épreuve, alors que compétences définies dans le plan d'études ; poids des différents domaines pas le même que dans le plan d'études, etc.), le poids ou la prise en compte de l'épreuve. D'autres remarques sont évoquées dans certains groupes mais pourraient être considérées comme générales : l'absence ou le manque de généralisation des prétests, la correction des EC réalisée par les enseignants, etc.

Globalement, on retrouve certaines critiques ou interrogations relevées dans une étude romande réalisée au primaire à Neuchâtel, Vaud et Fribourg (Diederonck, 2008) : moment de passation des épreuves, conditions de passation et de correction, utilisation des épreuves, qualité, notamment.

Au niveau de l'utilisation, on peut observer un certain consensus : la plupart des enseignants des deux niveaux les utilisent à des fins d'entrainement, voire de bachotage, ou encore à des fins de régulation de leur enseignement. Parfois, ils intègrent certains exercices dans leurs propres évaluations.

>>>

[22] Pour ne pas surcharger les élèves en particulier dans les petits degrés, il a été décidé par la direction de l'enseignement de ne pas tester tous les ans l'ensemble des domaines du français I (compréhension de l'écrit, compérhension orale et production écrite). Seule la compréhension de l'écrit est présente chaque année dans les trois degrés considérés, les deux autres sont présents à tour de rôle. Par ailleurs, chaque année, un texte d'un genre différent est sélectionné pour chaque degré.

Les propositions d'amélioration vont dans le sens des points faibles relevés. Ainsi, au *primaire*, elles concernent principalement la collaboration entre enseignants et concepteurs, le programme, les conditions de passation et de correction (pour garantir une plus grande objectivité, les classes devraient être croisées). Un certain nombre de suggestions se rapportent à une discipline : par exemple en français I, les mêmes composantes devraient être évaluées chaque année (et si possible toutes), la question de la production écrite est souvent mentionnée. En allemand, le niveau d'exigences est jugé trop facile. Enfin, le statut de l'évaluation en 2P est fréquemment questionné.

Les enseignants du CO évoquent également la question de la correction des travaux qui pourrait être croisée entre classes. Un des sujets les plus fréquemment cités par les enseignants des différentes disciplines est celui des différences de niveaux entre élèves et de leur prise en compte. Pour la plupart, l'épreuve ne devrait pas être commune (même en français, discipline enseignée en principe sans niveau). Pour d'autres, on pourrait avoir une partie commune et une autre uniquement pour les A et les niveaux forts. Concernant les barèmes, ils devraient être fixés a priori en fonction des attentes. De nombreuses remarques se réfèrent au type de connaissances ou de compétences évaluées ainsi qu'à leur répartition dans l'épreuve : dans plusieurs groupes de disciplines, les enseignants souhaiteraient une évaluation de compétences plus globales qui rendraient mieux compte de ce que savent les élèves, et non des connaissances pointues.

Les liens entre les résultats annuels des élèves et ceux à l'évaluation commune sont très variables d'un degré à l'autre, d'une école à l'autre, voire d'une discipline à l'autre. A l'école primaire, il semblerait qu'en mathématiques et en français II, les enseignants trouvent que les résultats aux EC confirment ceux de l'année. Pour les deux autres domaines, l'appréciation est nettement plus fluctuante.

Au CO, cela dépend de la discipline mais surtout du type de regroupement. Dans l'ensemble, les résultats à l'EC semblent confirmer ceux des évaluations de l'année pour la plupart des élèves de A et nettement moins pour ceux de B.

Enfin, les enseignants des deux niveaux d'enseignement font preuve d'un certain degré d'accord concernant les éléments positifs de l'évaluation cantonale ou commune. Par contre, pour ce qui concerne les éléments négatifs, le rejet est plus grand chez les enseignants de l'école primaire que chez ceux du CO.

4. Analyse de quelques épreuves de l'école primaire et du cycle d'orientation

4.1. Analyses docimologiques[23]

Cette partie du rapport vise à évaluer la qualité docimologique (essentiellement fiabilité, validité et stabilité) de quelques épreuves cantonales et communes de mathématiques et de français et s'inscrit dans les critères de qualité technique. Cette partie a été réalisée avec la précieuse collaboration de Daniel Bain (Groupe Édumétrie - Qualité de l'évaluation en éducation ; Société suisse pour la recherche en éducation, ancien chercheur en éducation au CRPP puis au SRED).

Nous nous sommes posé une première question fondamentale : les épreuves cantonales et communes déjà administrées étaient-elles fiables ? La fiabilité d'une épreuve dépendant de son utilisation, celle-ci sera évaluée selon plusieurs utilisations, certaines effectives, d'autres potentielles. Une utilisation effective des épreuves cantonales et communes est de participer, avec les notes de l'année, à la certification, c'est-à-dire de situer les résultats d'un élève sur l'échelle de connaissances/compétences ou de comparer les résultats d'un élève à un seuil de réussite. Une utilisation potentielle des épreuves serait de repérer les notions moins bien réussies qui mériteraient une révision au degré supérieur, ou de

[23] La docimologie est la science de l'évaluation en pédagogie.

hiérarchiser la difficulté des questions. Une troisième utilisation potentielle des épreuves pourrait être de classer des groupes d'élèves (selon leur classe ou leur établissement) sur la base des résultats aux épreuves. L'estimation de la fiabilité des épreuves pour ces trois types d'utilisation se fera au moyen de la théorie de la généralisabilité et occupera une grande place dans cette partie du rapport.

Nous nous sommes posé également la question de la validité (de contenu et prédictive) des épreuves. L'évaluation des acquis des élèves effectuée une année donnée peut-elle servir à estimer les chances qu'un élève suive avec plus ou moins de succès le programme auquel il sera confronté l'année suivante ? Nous chercherons aussi à savoir si des biais de contenu dans les épreuves peuvent affecter leur validité.

Nous nous focaliserons en troisième lieu sur la stabilité des épreuves avec l'exemple des mathématiques : la difficulté d'une épreuve est-elle stable d'une année sur l'autre pour un même degré ?

Enfin, nous proposerons quelques recommandations visant à améliorer la qualité technique des épreuves et nous mettrons en garde contre les risques qui surviennent lorsque certains critères de qualité technique ne sont pas respectés.

Épreuves cantonales à l'école primaire

Objectifs des épreuves

Selon les objectifs officiels (cf. partie II.1, p. 17), les épreuves cantonales participent, avec les notes de l'année, à la certification des élèves et au passage dans le degré ou le cycle suivant. Les épreuves de mathématiques et de français jouent également un rôle d'orientation par rapport aux regroupements par niveau en 7^{e}.

Les épreuves cantonales sont organisées à la fin de la 2P, de la 4P et de la 6P. Ces épreuves portent sur les disciplines suivantes : français I (comprenant la compréhension orale, la compréhension écrite et la production écrite), français II (appelé aussi « structuration » et comprenant la grammaire, l'orthographe, la conjugaison et le vocabulaire) et mathématiques, auxquelles s'ajoute l'allemand en 4P et 6P.

Les épreuves analysées ici peuvent être considérées comme des tests de maîtrise dans la mesure où elles satisfont aux conditions suivantes :

- les résultats d'un élève ne sont pas situés par rapport aux résultats d'un groupe mais comparés à un seuil de réussite,
- l'ensemble d'items/questions doit être suffisamment bien défini pour qu'on puisse donner une définition précise à la connaissance/compétence qui est observée,
- l'intérêt de l'épreuve est de savoir si la connaissance/compétence de l'élève est située au-dessus ou en dessous du seuil de réussite (définition adaptée de Cardinet et Tourneur, 1985).

Épreuves cantonales faisant l'objet d'une analyse

Suite à de précédents mandats, le SRED dispose de bases de données contenant les réponses des élèves aux épreuves cantonales de français et de mathématiques, pour un échantillonnage d'élèves présenté ci-après ainsi qu'au tableau 4.1.

Trois échantillons ont été utilisés dans ce rapport :

- échantillon d'élèves choisis au hasard, 2P et 6P, 2006 (appelé échantillon 1 dans ce rapport) ;
- échantillon d'élèves choisis de manière à représenter au mieux l'ensemble des élèves selon deux critères (catégorie socioprofessionnelle et langue parlée à la maison), 2P et 6P, 2007 (appelé échantillon 2) ;
- toutes les classes de sept écoles du Réseau d'enseignement prioritaire, 2P, 4P et 6P, 2007 (appelé échantillon REP).

Tableau 4.1. Épreuves cantonales faisant l'objet d'une analyse

Degré	Année	Population concernée	Épreuve
2P	2006	Échantillon 1 (n = 331)	Français
	2007	Échantillon 2 (n = 96)	Français
		Échantillon REP (n = 153)	Français + Mathématiques
4P	2007	Échantillon REP (n = 127)	Français + Mathématiques
6P	2006	Échantillon 1 (n = 381)	Français
	2007	Échantillon 2 (n = 94)	Français
		Échantillon REP (n = 150)	Français + Mathématiques

Fiabilité des épreuves pour l'évaluation des acquis des élèves

Méthode d'estimation de la fiabilité/fidélité[24] *des épreuves*

Ce paragraphe a pour objectif d'estimer la fiabilité de quelques épreuves cantonales (mathématiques et français) au moyen d'une analyse de généralisabilité (Bain et Pini, 1996)[25],[26]. En termes succincts, la fiabilité d'une épreuve désigne la capacité de l'épreuve à fournir des résultats répétables (c.-à-d. généralisables) et précis. Les résultats seront répétables (généralisables) si un élève obtient des résultats similaires dans d'autres conditions d'épreuves. Par exemple, est-ce qu'un élève obtiendrait les mêmes résultats aux épreuves si un autre enseignant avait corrigé l'épreuve ? Si au lieu de QCM, on avait utilisé une autre forme de questions ? Une épreuve sera généralisable si les résultats des élèves ne dépendent pas (ou de manière négligeable) des conditions d'épreuve, par exemple du choix du correcteur ou du répertoire d'items utilisé. Les résultats seront précis[27] si l'épreuve permet de distinguer de faibles différences dans les niveaux de réussite. Est-ce qu'un élève ayant obtenu 80% de réussite se situe au dessus du seuil de réussite fixé a priori (75% de réussite, par exemple) ? On ne pourra l'affirmer que si l'épreuve a une précision suffisante (c.-à-d. une faible marge d'incertitude). Avec une marge d'incertitude de 20% par exemple, on ne pourrait affirmer que cet élève est au-dessus du seuil de réussite. On aurait alors besoin d'une épreuve plus précise pour l'affirmer.

La fiabilité des épreuves peut être évaluée par trois coefficients de généralisabilité :

- un coefficient de généralisabilité *relatif* : évalue avec quelle confiance on peut classer des élèves, comme dans une évaluation à référence normative (type concours). Dans le cas particulier du dispositif d'évaluation étudié dans ce rapport (un ensemble d'items passés par des élèves), le coefficient de généralisabilité relatif correspond à l'alpha de Cronbach qui apprécie l'homogénéité (ou consistance inter-items) de l'épreuve. Ce coefficient relatif n'est pas utilisé dans ce rapport compte tenu de l'utilisation des épreuves cantonales et communes, qui n'est pas en principe de classer les élèves ;
- un coefficient de généralisabilité *absolu* : évalue avec quelle confiance on peut situer les résultats d'un élève sur l'échelle des connaissances/compétences, comme dans une évaluation à référence critériée ;

[24] Dans ce rapport, les termes *fiabilité* et *fidélité* sont utilisés comme synonyme l'un de l'autre.

[25] Une analyse de généralisabilité est basée sur l'analyse de variance et permet de différencier les différentes sources d'erreur de mesure.

[26] La fiabilité d'une épreuve est définie comme la proportion de variance du score « vrai » (moyenne qu'un sujet obtiendrait pour toutes les questions et conditions d'épreuves possibles) dans la variance totale des scores observés (moyenne obtenue par un élève pour une épreuve donnée qui correspond à un échantillon aléatoire d'items). Pour évaluer la fiabilité d'une épreuve, on estimera l'importance des différentes sources d'erreurs liées aux conditions de passation de l'épreuve (il s'agit ici de l'échantillonnage des élèves et des items).

[27] La précision est estimée à partir de l'écart type de l'erreur de mesure absolue.

- un coefficient de généralisabilité *absolu critérié*[28] qui prend en compte l'écart entre la réussite observée d'un élève et un seuil de réussite défini par les responsables de l'épreuve. Ce coefficient évalue avec quelle confiance on peut situer les résultats d'un élève par rapport à un seuil de réussite (ou en d'autres termes, différencier deux groupes d'élèves, ceux au-dessus et ceux en-dessous du seuil de réussite).

Les deux dernières approches permettent de décider si la maîtrise du domaine est acquise ou non pour un élève.

Ces coefficients varient de 0 à 1 et traduisent une fiabilité des épreuves d'autant plus élevée que leur valeur est proche de 1. Dans la pratique, en sciences de l'éducation, on considère généralement que la fiabilité d'une épreuve est satisfaisante lorsque la valeur des coefficients est au moins égale à 0.80[29].

Le modèle de généralisabilité nous a permis dans ce rapport d'estimer la fiabilité d'une épreuve. Il peut également fournir des indications pour l'améliorer. Il permet d'estimer combien d'items sont nécessaires pour atteindre une fiabilité satisfaisante et détermine si la suppression de certains items améliore la fiabilité de l'épreuve.

Résultats : évaluation de la fiabilité des épreuves pour l'évaluation des acquis des élèves

Les tableaux 4.2 et 4.3 rendent compte de la fiabilité (coefficients de généralisabilité absolu critérié et absolu) et de la marge d'incertitude pour les épreuves cantonales de français et mathématiques en classe primaire, passées en 2006 et 2007, d'une part pour les échantillons 1 et 2, d'autre part pour l'échantillon REP.

Intéressons-nous prioritairement à l'objet essentiel d'une épreuve à référence critériée, c'est-à-dire à l'écart entre la réussite d'un élève et le seuil de réussite fixé par les enseignants[30]. Cet écart peut-il être estimé fidèlement ? (voir colonne « Fiabilité par rapport à un seuil de réussite » du tableau 4.2). La réponse est « oui » pour l'essentiel des épreuves cantonales étudiées (cf. coef. ≥ 0.8). Celles-ci permettent pour la plupart de situer de façon suffisamment fiable les résultats des élèves par rapport au seuil de réussite fixé par les responsables des épreuves. En particulier, les épreuves de français I et II (structuration) ont une fiabilité satisfaisante pour cette utilisation. C'est le cas également des sous-parties du français I, hormis l'épreuve 2006 de compréhension orale de 2P (coef. = 0.66) et l'épreuve 2006 de compréhension écrite de 6P (coef. = 0.56). La première épreuve fera l'objet d'une analyse plus détaillée dans la suite de ce rapport (cf. paragraphe *complément d'analyse*). Quant à la seconde épreuve, son résultat est surprenant, d'autant plus que la fiabilité de l'épreuve de compréhension écrite pour le même degré passée l'année suivante (2007) a une fiabilité satisfaisante (coef. = 0.82). Le nombre plus élevé de questions a probablement contribué à améliorer significativement la fiabilité de cette épreuve.

La fiabilité des épreuves de français I et II pour situer les résultats des élèves sur l'échelle du test est un peu plus faible mais néanmoins proche de la limite considérée comme satisfaisante (cf. colonne « Fiabilité sur l'échelle du test » du tableau 4.2). En revanche, certaines sous-parties du français I posent problème pour cette utilisation. C'est le cas de l'épreuve de production écrite. La subjectivité de la correction (plus subjective que la correction de questions à choix multiples, par exemple) et le faible nombre de « questions » (plus exactement d'items d'évaluation) expliquent probablement en partie ce résultat.

Quant aux résultats basés sur l'échantillon REP, ceux-ci sont donnés à titre indicatif puisque les épreuves n'ont pas été conçues pour cet échantillon (qui est moins représentatif de la population d'élèves du canton que ne le sont les deux échantillons 1 et 2). Mentionnons néanmoins la fiabilité satisfaisante ou quasiment satisfaisante des épreuves de mathématiques en 2P et 6P.

---

[28] La théorie de la généralisabilité nomme ce coefficient Phi (lambda).

[29] Il existe d'autres seuils : l'APA *(American Psychological Association)* parle de seuil de suffisance lorsque l'alpha de Cronbach est supérieur à 0.7.

[30] Plus le seuil de réussite fixé par les responsables des épreuves sera en-dessous de la moyenne générale des élèves, plus le coefficient de généralisabilité absolu critérié sera élevé car on peut être sûr que la majorité des élèves atteindront ce seuil.

Tableau 4.2. Fiabilité des épreuves cantonales de français en classe primaire, échantillons 1 et 2, 2006 et 2007 (analyses de généralisabilité)

Degré	Épreuves	Année	Fiabilité* pour situer les résultats des élèves par rapport à un seuil de réussite	Fiabilité** pour situer les résultats des élèves sur l'échelle du test	Seuil de réussite (en %)	Marge d'incertitude (en %)	Nombre de questions de l'épreuve
2P	Français I***	2006	0.92	0.77	69	±11	35
	Compréhension écrite	2006	0.81	0.70	63	±20	12
		2007	0.81	0.71	68	±18	12
	Compréhension orale	2006	0.66	0.57	79	±17	14
	Production écrite	2006	0.93	0.44	67	±14	9
	Structuration****	2007	0.87	0.80	66	±16	11
6P	Français I	2006	0.87	0.70	68	±11	42
	Compréhension écrite	2006	0.56	0.55	68	±21	13
		2007	0.82	0.82	66	±16	20
	Compréhension orale	2006	0.71	0.47	68	±20	15
	Production écrite	2006	0.82	0.47	65	±18	14
	Structuration	2007	0.85	0.85	68	±13	17

*Estimée par le coefficient de généralisabilité absolu critérié.

**Estimée par le coefficient de généralisabilité absolu ; la fiabilité d'une épreuve est satisfaisante si le coefficient de généralisabilité est égal ou supérieur à 0.8.

***Les épreuves de compréhension orale, compréhension écrite et production écrite sont regroupées sous l'appellation 'français I'.

****L'épreuve de structuration comprend des questions de grammaire, orthographe, conjugaison et vocabulaire. Ces résultats ne sont pas pris en compte dans le bilan certificatif. Ils permettent d'apprécier les progrès accomplis en 'structuration' à la fin du cycle élémentaire et d'adapter l'enseignement au début du cycle moyen.

Tableau 4.3. Fiabilité des épreuves cantonales de français et mathématiques en classe primaire, échantillon REP, 2007 (analyses de généralisabilité)

Degré	Épreuves	Fiabilité pour situer les résultats des élèves par rapport à un seuil de réussite	Fiabilité pour situer les résultats des élèves sur l'échelle du test	Seuil de réussite (en %)	Marge d'incertitude (en %)	Nombre de questions de l'épreuve
2P	Compréhension écrite	0.65	0.63	68	±21	12
	Structuration	0.86	0.84	66	±17	11
	Mathématiques	0.77	0.72	60	±20	8
4P	Compréhension écrite	0.60	0.60	67	±21	16
	Structuration	0.76	0.70	68	±16	15
	Mathématiques	0.64	0.64	63	±22	12
6P	Compréhension écrite	0.78	0.76	66	±18	20
	Structuration	0.79	0.79	68	±14	17
	Mathématiques	0.84	0.84	67	±17	14

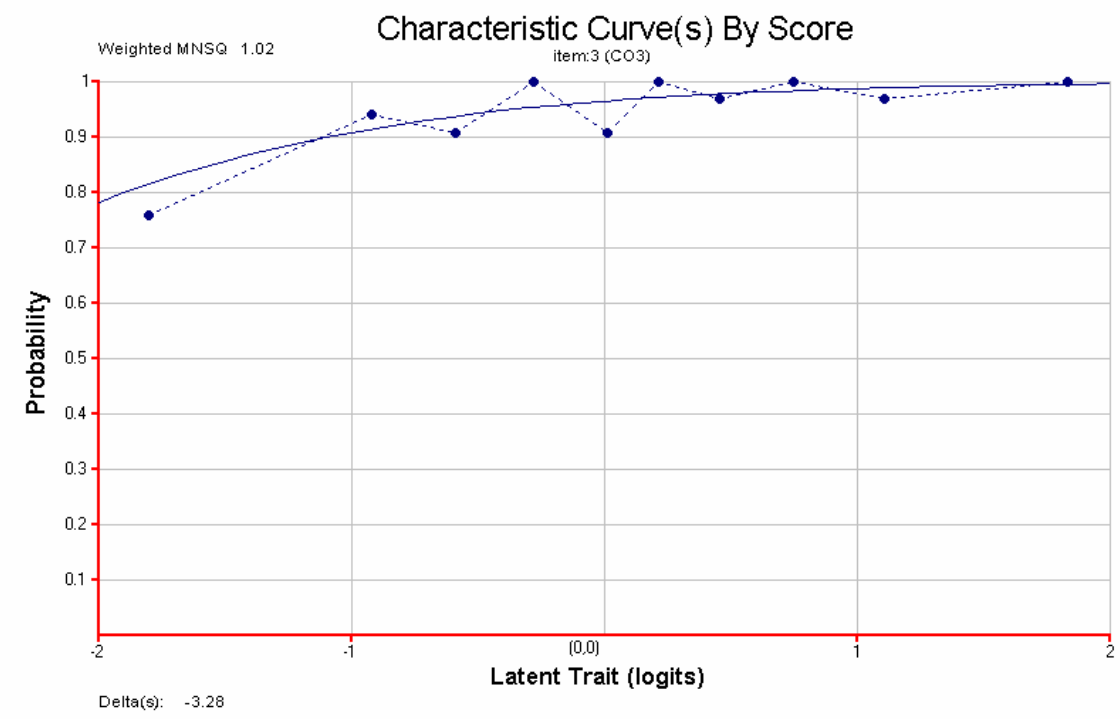
Complément d'analyse avec la théorie des réponses aux items[31] *de l'épreuve de compréhension orale de 2P (2006)*

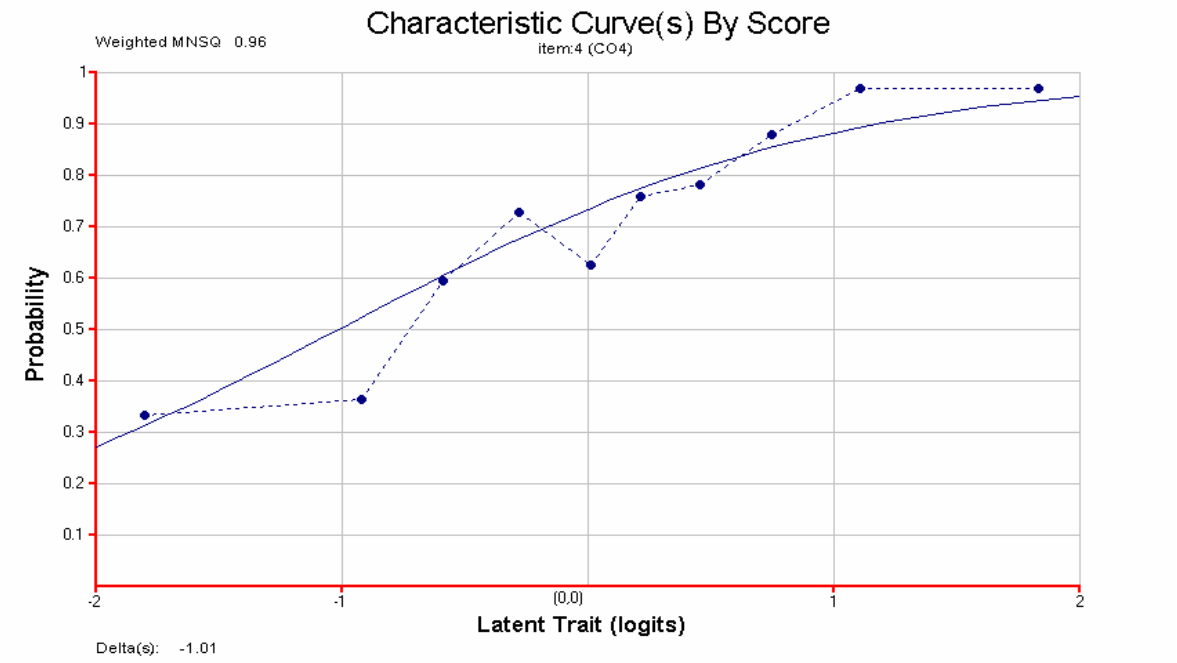
Nous utilisons ici la théorie des réponses aux items pour savoir pourquoi l'épreuve de compréhension orale de 2P administrée en 2006 a une fiabilité insuffisante (coefficient absolu égal à 0.57 ; tableau 4.2, colonne « Fiabilité sur l'échelle du test ») pour situer un élève sur l'échelle du test. L'analyse détaillée des questions de l'épreuve nous apprend que les taux de réussite aux items sont très élevés (10 questions sur 14 ont un seuil de réussite supérieur à 80%, 8 questions sur 14 ont un seuil de réussite supérieur à 90%). Nous illustrons en figures 4.1a et 4.1b deux courbes caractéristiques d'items (probabilité de réussir l'item en fonction des connaissances/compétences des élèves ; pour deux des quatorze items de l'épreuve). La probabilité de réussir le second item (figure 4.1b) dépend fortement de la connaissance/compétence des élèves ; ce qui n'est pas le cas du premier item (figure 4.1a) qui est réussi par pratiquement tous les élèves indépendamment de leur connaissance/compétence. Sur les quatorze items, dix ont une courbe caractéristique similaire à la première illustration (items peu discriminants) et quatre seulement ont une courbe s'approchant de la seconde illustration (qui correspond à une meilleure différenciation/discrimination des élèves).

De manière générale, une épreuve différencie ou discrimine mieux les élèves si elle ne comporte pas trop d'items très faciles ou très difficiles. Ces items « extrêmes » ne permettent en effet pas de répartir les élèves en deux groupes (ceux qui réussissent et ceux qui échouent) et ne contribuent pas à produire une variabilité suffisante des scores individuels (Pini et al., 2006, p. 28). Le choix d'items discriminants sera d'autant plus important dans les degrés où une orientation est nécessaire (6P et CO).

[31] La théorie des réponses aux items est un modèle statistique de la mesure relativement récent qui permet d'obtenir une fiabilité (appelée *information* dans ce contexte) et une précision qui dépendent du niveau des élèves. De plus, ce type de modèle estime conjointement la difficulté des items (entre autres) et la compétence d'un élève. Ainsi, un élève ayant réussi, par exemple, 50% d'items difficiles aura une compétence (estimée par le modèle) plus élevée qu'un élève ayant réussi 50% d'items faciles.

Figures 4.1a et b. Courbes caractéristiques de deux items de l'épreuve de compréhension orale de 2P (2006)



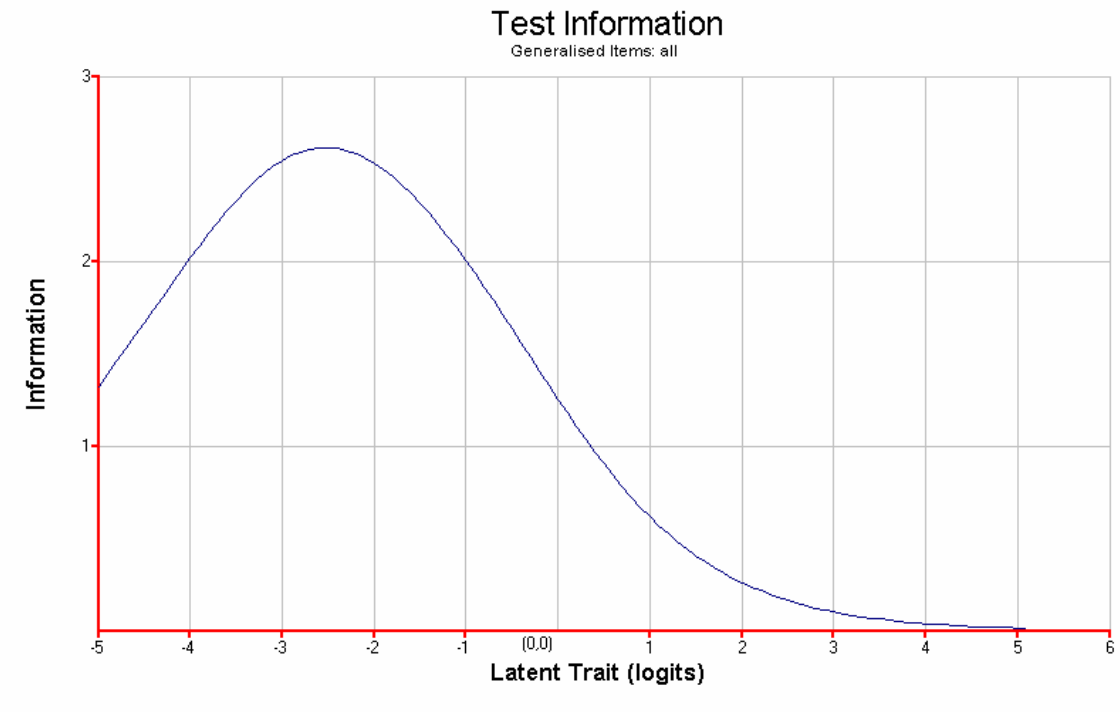


Légende

Probabilité de réussir l'item en fonction des connaissances/compétences des élèves *('Latent Trait')* :

- probabilité mesurée = trait pointillé
- probabilité estimée = trait continu

Figure 4.2. Courbe d'information de l'épreuve de compréhension orale de 2P (2006)



La courbe d'information de l'épreuve (cf. figure 4.2) illustre dans quelle échelle de connaissances/compétences l'épreuve est la plus fiable (c.-à-d. la plus informative). Ici, l'épreuve permet seulement de distinguer les connaissances/compétences des élèves les plus faibles, ceux ayant une connaissance/compétence (appelés ici « latent trait ») de (–)2.5, ce qui correspond à des élèves qui maîtriseraient 50% du contenu de l'épreuve ; c'est le cas pour cette épreuve de 10 élèves sur les 325 échantillonnés ! En revanche, l'épreuve de compréhension orale ne permet pas bien de distinguer les connaissances/compétences des autres élèves (et donc de les noter). Idéalement, l'épreuve devrait être fiable sur une échelle de connaissances/compétences beaucoup plus étendue et notamment là où l'on souhaite différencier les élèves, c'est-à-dire autour du seuil de réussite (79% de réussite, soit une compétence de –0.8 dans le cas particulier de cette épreuve).

Fiabilité des épreuves pour repérer les notions/questions difficiles

Les épreuves pourraient être utilisées pour comparer entre elles la difficulté de chaque question (hiérarchiser la difficulté des questions) ou pour identifier les questions qui posent problème (celles, par exemple, qui seraient réussies par moins de la moitié des élèves). Peut-on dire, par exemple, qu'une question réussie par 60% des élèves est plus facile qu'une autre présentant un taux de réussite de 50% ? On pourra l'affirmer uniquement si l'épreuve a une fiabilité suffisante pour cette utilisation (classement de la difficulté des items).

Pour cette utilisation-là, les épreuves de mathématiques et français analysées dans ce rapport ont une excellente fiabilité (tableau de résultats non présenté dans ce rapport). Ainsi, ces épreuves déjà administrées de mathématiques et français auraient pu tout à fait être utilisées pour repérer les notions moins bien réussies qui auraient mérité une révision au degré supérieur ou pour hiérarchiser le degré de difficulté des items.

Validité des interprétations des scores aux épreuves cantonales

Suivant la conception de la validité exposée dans l'ouvrage de Bertrand et Blais (2004), nous considérons ici plusieurs stratégies pour valider les interprétations faites à partir des scores à une épreuve.

Validité de contenu

Nous n'abordons que très succinctement dans ce rapport la validité de contenu que nous supposons garantie par l'expertise que les auteurs des épreuves ont du programme en mathématiques et français des divers degrés du primaire. Rappelons que la validité de contenu est assurée par un échantillon aussi représentatif que possible des objectifs qui composent le domaine de référence.

Le guide méthodologique pour l'évaluation et la mesure en éducation de Pini et al. (2006) met en garde contre « la décomposition de processus, par nature, complexes, en une série d'activités élémentaires qui risque d'aboutir à une atomisation des compétences testées ». L'évaluation ne permettrait alors pas de déterminer si l'élève est réellement capable d'articuler et de coordonner ses connaissances et compétences, ce qui mettrait en jeu la validité de l'épreuve. Il semble que l'épreuve de production écrite soit concernée par cette mise en garde : un élève qui respecte tous les « items d'évaluation » aura une bonne note mais n'a pas forcément produit un texte jugé satisfaisant par l'enseignant (cf. partie 3, *Le point de vue des enseignants*).

Validité prédictive d'une épreuve

Dans ce paragraphe, nous souhaitons vérifier si l'évaluation externe effectuée une année donnée peut servir à estimer les chances qu'un élève suive avec plus ou moins de succès le programme auquel il sera confronté l'année suivante. Dans ce cas, le score aux épreuves cantonales obtenu dans un domaine une année donnée devrait être corrélé significativement et fortement au score obtenu dans ce même domaine l'année suivante.

Compte tenu des données disponibles, ce type de prédiction peut être réalisée pour l'épreuve de mathématiques de 6P administrée en 2007. La validité prédictive des épreuves cantonales est d'autant plus importante à évaluer en 6P que les épreuves cantonales servent, avec les notes de l'année, à orienter les élèves dans des regroupements.

Tableau 4.4. Corrélation entre le score en mathématiques obtenu en 6P (2007) et celui obtenu en 7e (2008), par regroupement

Regroupement en 7e	Nombre d'élèves	Corrélation
A	2534	0.58
B + C	613	0.49
H	575	0.72

Au tableau 4.4 sont présentées les corrélations entre le score en mathématiques obtenu en 6P (2007) et celui obtenu en 7e (2008), par regroupement. L'information donnée par le score en mathématiques en 6P a une valeur prédictive appréciable des futures performances en mathématiques à l'épreuve commune de 7e, pour les classes hétérogènes[32]. La corrélation de 0.72, élevée au carré (0.52), nous donne la part de variance expliquée par le modèle. En d'autres termes, le score en mathématique en 6P permet de prédire 52% de la variance des scores obtenus en mathématiques en 7e. Notons d'autant plus que le changement de type d'enseignement (du primaire au CO), ainsi que le poids relativement faible de l'épreuve de mathématiques comparé aux notes de l'année affectent cette relation entre les résultats aux épreuves pour les deux degrés successifs.

La gamme plus large de résultats obtenus dans les classes hétérogènes par rapport aux classes de niveau A (absence d'élèves très faibles) ou de niveau B ou C explique partiellement pourquoi la validité prédictive de l'épreuve de mathématiques en 6P est meilleure en classes hétérogènes que pour les niveaux A ou B + C.

[32] La validité prédictive est habituellement considérée satisfaisante à partir d'une corrélation de 0.7 entre les résultats de degrés successifs (Moody, 2001).

Facteurs pouvant affecter la validité d'une épreuve

Nous évaluons ici un biais possible qui pourrait se formuler de la manière suivante : « les épreuves de mathématiques font appel aussi à des connaissances/compétences en compréhension écrite », ce qui ajouterait à l'épreuve de mathématiques une autre dimension de compétence. Est-ce que les élèves bons en compréhension écrite réussissent mieux les questions de mathématiques à énoncés plus longs et plus complexes que les questions de mathématiques à énoncés courts ? Si oui, cela indiquerait un biais de validité puisque l'épreuve de mathématiques mesurerait des connaissances/compétences non seulement en mathématiques mais aussi en compréhension écrite.

Les corrélations entre les scores de mathématiques et de compréhension écrite ont été étudiées en 6P pour l'échantillon du REP[33] (2007). La relation entre les scores de compréhension écrite et ceux de mathématiques est de même ampleur que l'on considère les questions de mathématiques à énoncés relativement plus longs et complexes ou les questions de mathématiques à énoncés courts et de formulation simple. Il semble donc que les élèves plus faibles en compréhension écrite ne soient pas désavantagés pour la résolution de problèmes de mathématiques à énoncés longs. On peut par conséquent exclure l'existence d'un biais qui aurait affecté la validité de l'épreuve de mathématiques.

Épreuves communes du cycle d'orientation

Objectifs des épreuves

Selon les objectifs officiels (cf. partie II.1), les épreuves communes participent à la certification des élèves et au passage dans le degré ou le cycle suivant.

Épreuves communes faisant l'objet d'une analyse

Nous avions à notre disposition les épreuves de mathématiques de 7e et 9e (2008) pour l'ensemble des élèves et l'épreuve de français (2007) pour l'échantillon 2 (échantillon d'élèves choisis de manière à représenter au mieux l'ensemble des élèves selon deux critères : catégorie socioprofessionnelle et langue parlée à la maison ; cf. tableau 4.5).

Tableau 4.5. Épreuves communes faisant l'objet d'une analyse

Degré	Année	Population concernée	Épreuve
7e	2008	Ensemble des élèves	Mathématiques
9e	2007	Échantillon 2 (n = 134)	Français
	2008	Ensemble des élèves	Mathématiques (2 épreuves)

Fiabilité des épreuves pour l'évaluation des acquis des élèves

Les épreuves de mathématiques de 7e et 9e (2008) et celle de français de 9e (2007) ont une fiabilité satisfaisante, que l'on cherche à situer les élèves sur une échelle de connaissances/compétences ou à évaluer leur niveau de maîtrise en le comparant à un seuil fixé (coef. de généralisabilité proches ou supérieurs à 0.8 ; cf. tableau 4.6). Pour information, la fiabilité de l'épreuve de tronc commun en 9e de 2008 est similaire à celle de 2006, voire légèrement supérieure (cf. Bain, Weiss, Agudelo (à paraître)). En revanche, même si l'épreuve de français est fiable pour ces deux utilisations, ses sous-parties (en particulier, langue et moyens langagiers) évaluent des domaines de connaissances/compétences qui n'ont pas une fiabilité suffisante pour situer les résultats des élèves sur l'échelle du test. Les scores obtenus dans ces différents domaines ne devraient donc avoir qu'une valeur indicative.

[33] L'échantillon du REP est le seul échantillon permettant l'analyse conjointe des épreuves de mathématiques et de français.

Mentionnons pour information que la manière de répondre des élèves aux épreuves de mathématiques indique une seule dimension prépondérante[34] alors que ces épreuves sont censées mesurer plusieurs domaines.

Tableau 4.6. Fiabilité des épreuves communes de mathématiques et français du CO (analyses de généralisabilité)

Degré	Épreuves	Année	Regroupement	Fiabilité pour situer les résultats des élèves par rapport à un seuil de réussite	Fiabilité pour situer les résultats des élèves sur l'échelle du test	Marge d'incertitude (en %)	Nombre de questions de l'épreuve
7e	Math	2008	H	0.90	0.89	±14	18
			A	0.87	0.85	±14	18
			B+C	≥ 0.78	0.78	±15	18
9e	Math	2008	Tous (tronc commun)	≥ 0.92	0.92	±13	14
			Seconde épreuve	≥ 0.85	0.85	±16	12
	Français	2007	Tous	≥ 0.83*	0.83*	±10	53
	Contenu			- **	0.73	±16	18
	Langue			- **	0.62	±18	21
	Moyens langagiers			- **	0.60	±18	14

N.B. Le seuil de réussite n'est pas indiqué dans ce tableau car il diffère selon les regroupements.

* Valeur probablement surestimée car non-indépendance des questions (une même question permet d'attribuer des points à différents domaines).

** Le seuil de réussite est fixé pour l'ensemble de l'épreuve de français et non par domaine.

Fiabilité des épreuves pour repérer les notions/questions difficiles

Comme pour les épreuves du primaire, les épreuves de mathématiques (2008) et de français (2007) du CO ont une excellente fiabilité pour repérer les notions moins bien réussies qui mériteraient une révision au degré supérieur ou pour hiérarchiser le degré de difficulté des items.

Fiabilité des épreuves pour établir un classement des classes au sein d'un établissement

En 2008, les établissements auraient-ils pu utiliser les épreuves communes de mathématiques pour établir un classement de leurs classes de 7e et ainsi proposer éventuellement des mesures d'accompagnement aux classes les plus faibles ? Posons-nous cette question pour l'épreuve commune administrée aux élèves du regroupement A.

Dans ce cas présent où l'on s'intéresse à la fiabilité des épreuves pour établir un classement des classes, plusieurs sources de variation vont nuire à l'estimation du niveau moyen des classes d'un établissement. Citons les deux principales sources de variation influant sur les résultats :

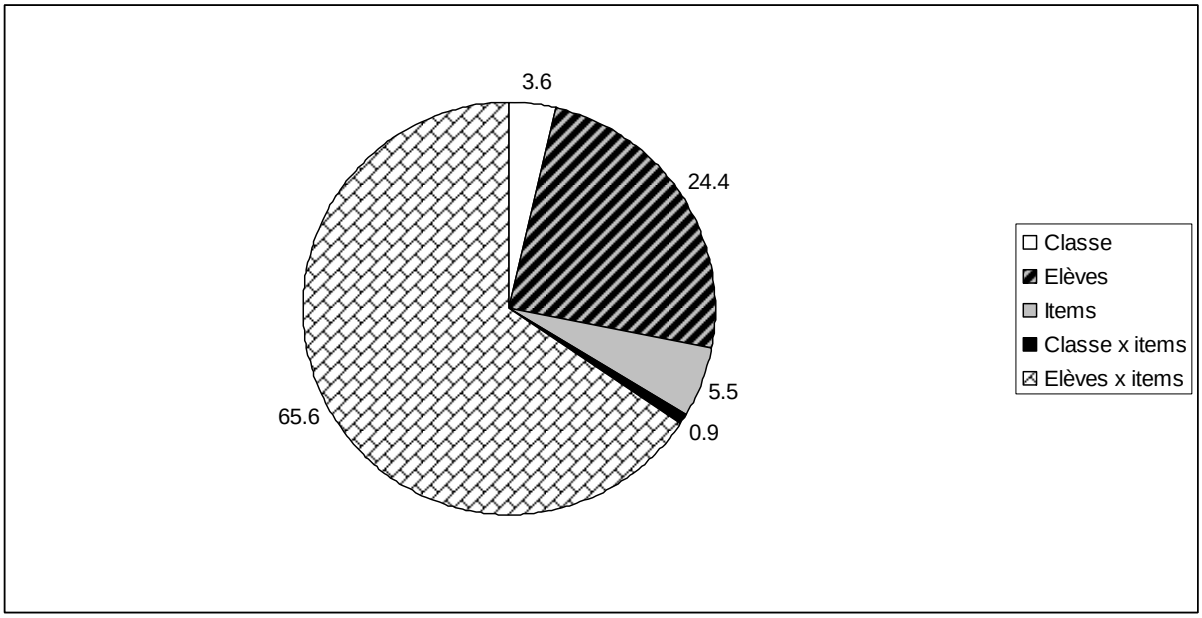
- les différences de réussite entre élèves d'une même classe. Notons que les différences entre les élèves d'une même classe sont en général beaucoup plus marquées que les différences entre les classes[35] ;

[34] Mise en évidence par une analyse factorielle.

- l'interaction entre les élèves et les items : les élèves réagissent différemment aux items indépendamment de leur niveau moyen de compétence et du niveau de difficulté des questions. Le classement des élèves dépend de l'item considéré : certains élèves forts échouent parfois à des items faciles, peut-être parce qu'ils voient des pièges là où il n'y en a pas ; inversement, des élèves plus faibles donnent une bonne réponse à des questions généralement mal réussies, peut-être aidés par le hasard (dans le cas des réponses à choix multiples) (cf. Bain et Pini, 1996).

La figure 4.3 illustre l'importance des diverses sources de variation lorsqu'on souhaite comparer le niveau moyen des classes d'un établissement (ici l'exemple porte sur un établissement choisi au hasard). Précisons que cette décomposition de la variation est tout à fait représentative de ce qui se passe dans les autres établissements.

Figure 4.3. Importance (en %) des différentes sources de variation intervenant dans l'estimation du niveau moyen des classes d'un établissement



Les différences de réussite entre classes sont faibles (ici elles comptent pour 3.6% de la variation totale ; figure 4.3) par rapport à d'autres sources de variation (différences entre élèves, interaction entre élèves et items)[36]. Ces dernières introduisent un flou non négligeable dans l'évaluation des différences entre classes. Pour la majeure partie des établissements, l'épreuve de mathématiques n'a pas une fiabilité suffisante pour distinguer les différences entre classes d'un établissement[37].

En supposant que l'épreuve ait une fiabilité suffisante pour distinguer les différences entre les classes d'un établissement, celles-ci devraient être interprétées avec prudence tant que l'on n'est pas assuré

[35] Par exemple, la variance intra-classe est dix fois supérieure à la variance inter-classes dans le cas du français en 6P (cf. Soussi et al., 2008 ; p. 156).

[36] Soussi et al. (2008) reportent une variabilité inter-classes légèrement plus élevée qui confirme des constats effectués dans d'autres études sur les effets-classes (Bressoux, 1994). En dehors du fait qu'il s'agit de disciplines différentes (français dans l'étude réalisée par Soussi et al. *vs* mathématiques dans le cas présent), de degrés (6P *vs* 7e) et de modèles différents (multi-niveaux *vs* analyse de généralisabilité), l'analyse porte ici sur les seules classes au sein d'un établissement (puisqu'on se place dans le cas d'un établissement qui souhaite comparer ses classes entre elles, et seulement elles) et non sur l'ensemble des classes du canton (pouvoir statistique alors beaucoup plus élevé pour évaluer l'effet inter-classes). En revanche, le modèle de généralisabilité a l'avantage de tenir compte de la variabilité de réponse aux items.

[37] Coefficients de généralisabilité calculés par établissement pratiquement tous inférieurs à 0.8 (sauf pour un établissement) ; résultats non présentés dans ce rapport.

que les conditions de passation et/ou de correction sont homogènes d'une classe à l'autre. Nos propos tiennent ici à mettre en garde contre l'utilisation des épreuves communes comme moyen systématique d'identifier des différences entre classes ou entre établissements. Il va de soi qu'une comparaison des établissements sur la base des épreuves communes nécessite également et impérativement une vérification similaire à celle présentée ici.

De plus, notons qu'un tel classement sur la base des épreuves communes de 2008 ne prend en considération qu'une partie des directives en matière d'enseignement. La réussite des études en général est plus complexe que les seuls résultats à une épreuve qui n'est d'ailleurs pas conçue spécifiquement pour un tel objectif de classement des classes ou des établissements.

Validité des interprétations des scores aux épreuves communes

Les analyses réalisées pour le primaire – validité prédictive et identification de biais pour l'épreuve de mathématiques – n'ont pas pu être appliquées aux épreuves du CO faute de données disponibles.

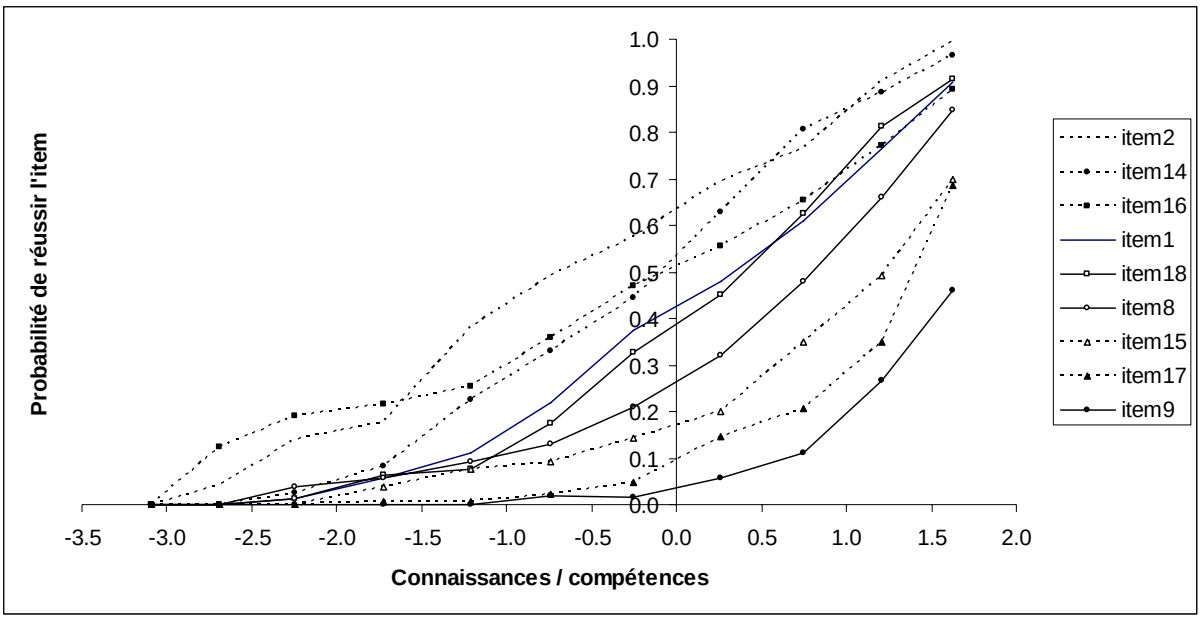
Stabilité des épreuves : difficulté des épreuves de mathématiques de 7e A, 2007 et 2008

La comparaison des volées d'écoles de même degré sera grandement facilitée si les épreuves ont des caractéristiques similaires d'une année à l'autre (même difficulté, par exemple).

Nous comparons ici la difficulté des deux épreuves de mathématiques de 7e A, l'une administrée en 2007 et l'autre en 2008. Pour cette comparaison, nous recourrons à nouveau à la théorie des réponses aux items.

Chaque item d'une épreuve peut être caractérisé par une courbe (appelée courbe caractéristique d'item) représentant la probabilité de réussir l'item en fonction des compétences/connaissances des élèves (estimées sur la figure 4.4 par le score global standardisé de mathématiques).

Figure 4.4. Courbes caractéristiques de quelques items de l'épreuve de mathématiques, 7e A, 2008

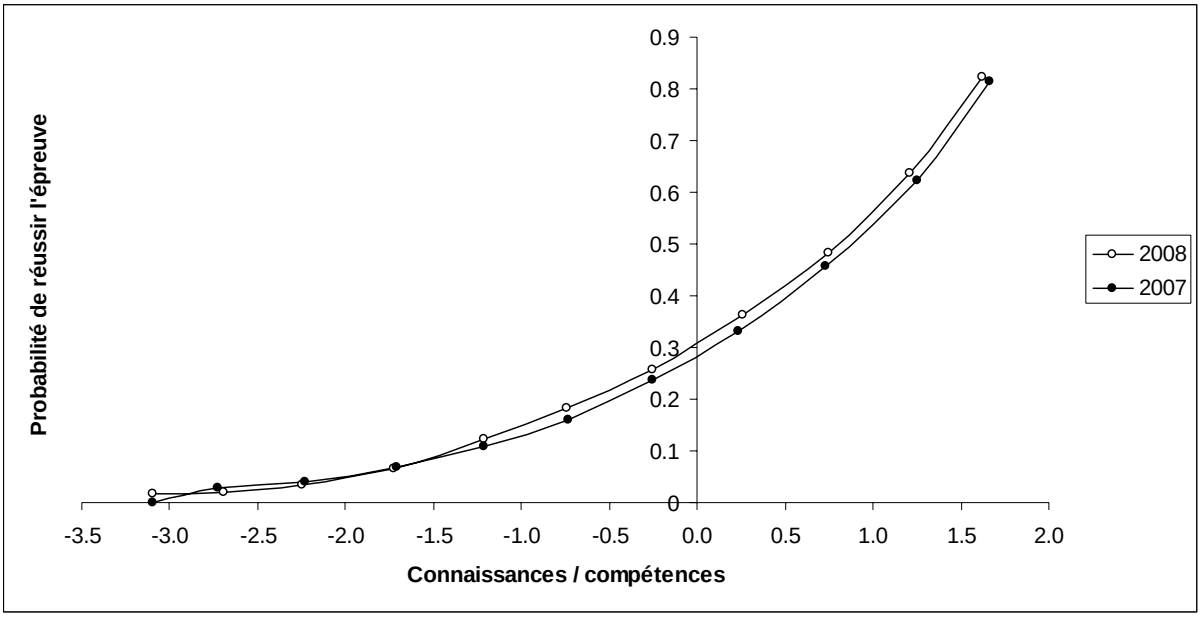


L'item le plus en bas de la figure est un item difficile (item 9 ; la probabilité de le réussir est faible), celui le plus en haut de la figure est plus facile (item 2 ; la probabilité de le réussir est forte).

La courbe caractéristique du test (c.-à-d. de l'épreuve) s'obtient en additionnant chacune des courbes caractéristiques des items de l'épreuve. Il s'agit en fait de faire la somme des probabilités à chaque

niveau de compétences/connaissances. La figure 4.5 représente la courbe caractéristique de test pour l'épreuve de mathématiques administrée en 2007 et celle administrée en 2008[38].

Figure 4.5. Courbes caractéristiques de test : épreuve de mathématiques, 7e A, 2007 et 2008



Comme pour une courbe caractéristique d'item, la courbe caractéristique de test donne, graphiquement, une idée de la difficulté globale de l'épreuve. Ici, la courbe de l'épreuve de 2008 est très proche de celle de 2007, indiquant que les deux épreuves ont été de difficulté très similaire.

Propositions d'une démarche d'amélioration de la qualité technique des épreuves et erreurs à éviter

Les auteurs de ce rapport ont présenté un panel d'analyses sur des épreuves déjà administrées qu'il conviendrait de mener à toute nouvelle session. Ceci afin de déterminer si l'épreuve nouvellement administrée répond à un objectif donné de manière fiable ou non. La fiabilité d'une épreuve pour son objectif primordial, la certification, devrait être évaluée lors des pré-tests. Les concepteurs pourraient ainsi améliorer si besoin les épreuves afin qu'elles remplissent au mieux leur rôle essentiel de certification. Un pré-test est également nécessaire pour fixer les seuils de réussite d'une épreuve. Cela permet de confronter les seuils fixés a priori aux résultats effectivement obtenus lors du pré-test (cf. Pini et al., 2006 ; p. 25). Cela peut être aussi l'occasion de pré-tester l'épreuve avec des élèves de différents niveaux et de vérifier si l'épreuve permet de mettre en évidence ces différences de niveaux par différents taux de réussite (Pini et al., 2006 ; p. 29). Quant à d'autres utilisations potentielles des épreuves (identifier les items moins bien réussis pour adapter l'enseignement l'année suivante, comparer des classes qui n'auraient pas suivi la même méthode d'enseignement), il est impératif de vérifier si les épreuves ont une fiabilité suffisante pour ces utilisations.

Pour les épreuves futures, il conviendrait de constituer progressivement une banque d'items répertoriant les items selon leur niveau de difficulté, leur pouvoir de discrimination, leur format (cf. Pini et al., 2006).

[38] Une courbe caractéristique de test représente en ordonnée la somme des probabilités de réussir chaque item. L'ordonnée varie donc entre 0 et *n*, avec *n* le nombre d'items de l'épreuve (16 items en 2007, 18 items en 2008). Pour des raisons de comparabilité des deux épreuves, l'ordonnée de la figure 4.5 représente ici la somme des probabilités divisée par le nombre d'items de l'épreuve.

Il serait ainsi plus aisé de constituer une épreuve qui réponde aux objectifs fixés (p. ex. être de même difficulté que la session précédente, différencier les élèves pour les orienter au mieux).

En outre, plusieurs éléments peuvent affecter la fiabilité d'une épreuve, notamment la correction par les enseignants. Si, par exemple, plusieurs enseignants n'attribuent pas les mêmes scores aux mêmes copies, certains favorisant ou défavorisant sciemment ou inconsciemment un élève qu'ils connaissent, la fiabilité de l'épreuve peut être mise à mal. En revanche, une correction en 'aveugle' (sans connaître l'identité de l'élève) ou en changeant l'ordre des copies lorsque plusieurs correcteurs interviennent améliorera la fiabilité de l'épreuve (Pini et al., 2006).

Résumons les points principaux abordés dans cette partie du rapport par une liste d'erreurs à éviter (cf. Gullickson, 2002) :

- supposer que les épreuves peuvent être utilisées pour la certification et l'orientation des élèves sans avoir évalué la fiabilité des épreuves ;
- supposer qu'une épreuve, si elle est fiable pour une utilisation (p. ex. la certification), l'est forcément pour une autre utilisation (p. ex. le classement des établissements) ;
- supposer qu'une épreuve, si elle est fiable pour une population (p. ex. regroupement A), l'est également pour une autre population (p. ex. regroupement B) ;
- supposer qu'une épreuve est fiable quelles que soient les conditions de passation ou de correction. Celles-ci peuvent grandement affecter la fiabilité d'une épreuve ;
- ne pas tenir compte de la marge d'incertitude pour la certification d'un élève dont les résultats sont proches du seuil de réussite ;
- supposer que la fiabilité d'une épreuve garantit sa validité.

Tableau 4.7. Synthèse des analyses docimologiques[39]

	École primaire	CO
POINTS FORTS		
Qualité technique		
Fiabilité des épreuves (français, mathématiques) pour l'évaluation des acquis des élèves (certification)	2006, 2007 (échantillon d'élèves)	2007, 2008
Fiabilité des épreuves (français, mathématiques) pour repérer les notions qui posent problème (régulation de l'enseignement)	2006, 2007 (échantillon d'élèves)	2007, 2008
Stabilité du point de vue de la difficulté des épreuves (mathématiques)	Données non disponibles	2007, 2008
Efficience/efficacité		
Validité prédictive de 6P à 7e	2007 à 2008	

POINTS FAIBLES		
Qualité technique		
Fiabilité (situer un élève sur l'échelle du test) insuffisante pour évaluer des connaissances/compétences détaillées (p. ex. au primaire : compréhension orale en 2P, production écrite ; p. ex. au cycle : langue et moyens langagiers) *	2006, 2007 (échantillon d'élèves)	2007, 2008
Fiabilité insuffisante (mathématiques) pour classer les classes d'un établissement	Données non disponibles	2008

* Cela implique que l'évaluation des connaissances/compétences détaillées ne doit être utilisée que de manière indicative.

[39] Cf. page précédente pour les propositions d'une démarche d'amélioration de la qualité technique des épreuves et les erreurs à éviter.

4.2. Comparaison d'épreuves de 6P et de 7e en français et mathématiques : l'exemple des EC 2007-2008

Il nous a semblé intéressant de compléter les analyses docimologiques par un point de vue qualitatif en comparant les approches respectives utilisées pour élaborer les épreuves dans les deux ordres d'enseignement dans des domaines communs, les mathématiques et le français, en se centrant sur deux exemples en 6P et en 7e.

Évaluation du français en 6P et 7e

En ce qui concerne le français, nous n'entrerons pas dans le détail du contenu des plans d'études de 6P et de 7e étant donné leur grande complexité et leurs différences. A l'école primaire, dans le plan d'études *(Les objectifs d'apprentissage de l'école primaire)*, l'enseignement du français est organisé en deux principaux domaines : comprendre et produire des textes de genres différents à l'écrit et à l'oral et, observer le fonctionnement de la langue. Le premier s'articule autour de quatre composantes : le contexte (s'adapter aux situations de communication), le contenu (élaborer le thème et les idées), la planification (dégager l'organisation du texte) et la textualisation (utiliser les ressources de la langue). Le second se compose de la grammaire, le vocabulaire, l'orthographe et la conjugaison.

L'enseignement du français au CO comporte plusieurs champs d'études : le domaine de la langue (vocabulaire, syntaxe, conjugaison, orthographe), celui des discours (moyens permettant la construction des textes et des discours, celui des textes (types et genres) ainsi que celui de la méthodologie (dimension transdisciplinaire du français). L'évaluation de 7e comprend la lecture d'un texte narratif et des questions de langue. Nous allons maintenant décrire les épreuves.

De manière globale, l'approche est assez différente. En 6P (comme pour les deux autres degrés pris en compte au primaire), elle est organisée par domaines : en 2008, trois domaines ont été pris en compte dans l'épreuve de 6P en français : la compréhension de l'écrit et la production écrite pour le français I ainsi que la structuration ou français II.

L'épreuve de 7e est organisée de manière davantage intégrée avec une première partie consacrée à la compréhension et une deuxième partie à la grammaire et l'expression écrite. Pour chacune de ses parties, les questions peuvent relever de plusieurs composantes (contenu, langue et moyens langagiers). Les résultats sont également déclinés selon ces trois composantes.

Voici plus précisément comment les différentes épreuves sont organisées.

Épreuve de 6P

A l'école primaire, les épreuves comportent une table de spécification qui permet de renvoyer pour chaque question à des objectifs.

Tableau 4.8. Composition de l'épreuve de français de 6P

Domaines	Durée	Nombre d'items	Seuil de réussite
Compréhension d'un texte écrit (récit)	1h30	18	16/24pts
Production écrite (Écoute d'une émission en vue de la production d'un texte explicatif sur le fonctionnement d'un ballon à air chaud)	2h15	6	14/22 pts
Structuration (grammaire, orthographe, conjugaison, vocabulaire)	1h30	17	45/69pts

De manière générale, les formats des questions sont variés : QCM, questions vrai/faux, questions à réponse courte, questions ouvertes.

L'épreuve comporte une unité thématique (par exemple en 2007-2008 : le voyage en ballon). La production écrite avec sa mise en contexte (écoute d'une émission radiophonique) constitue une véritable séquence didactique.

Pour les deux domaines du français I (compréhension d'un texte écrit et production écrite), les différentes questions se distribuent de la manière suivante.

Tableau 4.9a. Composition à l'intérieur des deux domaines du français I

Français I	*Composantes*	Objectifs	Nombre de questions	Nombre de points
Compréhension de l'écrit	*Contexte*	Identifier le but du texte	1	1
	Contenu	- Questionner le texte sur les contenus propres au genre - Dégager l'idée véhiculée par le titre - Repérer dans un texte l'élément d'information recherché - Comprendre une information implicite - Situer dans le texte une information implicite	7	9
	Planification	- Avoir une représentation de l'ensemble de la narration - Comprendre les relations avec les différentes parties du texte	2	2
	Textualisation	- Comprendre le sens d'un mot, d'une phrase - Interpréter les différents temps du verbe - Interpréter les reprises pronominales et nominales - Comprendre une paraphrase	8	12
Production écrite	*Contexte*	- Adapter sa production en fonction de la situation de communication	1	2
	Contenu	- Comprendre les données ou les mécanismes qu'on veut présenter ou expliquer - Donner des informations vraies	3	11
	Planification	- Donner un titre ou un sous-titre qui résume l'information véhiculée par le texte - Respecter l'ordre chronologique, établir des relations de causes	2	3
	Textualisation	- Respecter les contraintes orthographiques et syntaxiques	2	6

Pour le français II ou structuration, l'épreuve est constituée de la manière suivante.

Tableau 4.9b. Composition à l'intérieur du domaine *structuration* du français II

	Composantes	Objectifs	Nombre de questions	Nombre de points
Structuration	*Grammaire*	Connaître les fonctions grammaticales Connaître les catégories grammaticales Connaître les types et formes de phrases Connaître les valeurs sémantiques	6	21 (14)
	Orthographe	Connaître l'orthographe des mots-outils Connaître les accords dans le GN Accorder le verbe avec son sujet Copier sans faute Respecter l'orthographe de ses propres productions	4	19 (12)
	Conjugaison	Connaître les verbes de la liste du plan d'études	3	12 (8)
	Vocabulaire	Produire des mots d'une même famille Former des mots à l'aide de la dérivation Connaitre les relations entre les mots Reconnaître différents sens d'un même mot Utiliser un dictionnaire	4	17 (11)

Notons que pour chaque sous-domaine de la structuration, un seuil de réussite a été fixé (il figure entre parenthèses).

Épreuve de 7e

Il faut d'abord préciser que contrairement à d'autres disciplines ou d'autres degrés, l'épreuve de français de 7e est commune à tous les regroupements. Ce n'est que le barème qui est adapté au type d'élèves (A et H, B et C). Ce barème est déterminé a posteriori sur la base des résultats des élèves, le seuil de réussite (c'est-à-dire la note de 3.5) se situant autour de 60-70% des points pour les élèves de A et de H et autour de 50-55% pour les élèves de B et de C. Dans l'épreuve, il n'y a pas de table de spécification, la répartition se fait selon trois composantes : contenu, langue et moyens langagiers. Il est prévu que le nombre de points doit être comparable dans ces trois composantes. Il est toutefois précisé dans un document élaboré l'année passée et qui devrait servir en 2008-09 qu'en cohérence avec le plan d'études, un accent est porté sur l'une ou l'autre dimension : par exemple en 7e, il y aura davantage de points pour la Langue. Dans un premier temps, nous allons présenter la structure générale de l'épreuve puis essayerons de répartir les questions en les attribuant à des domaines équivalents de ceux de la 6P afin de pouvoir les comparer.

La durée totale de l'épreuve est de 95 mn.

Tableau 4.10. Composition de l'épreuve de français de 7e

Composantes Domaines	Contenu	Moyens langagiers	Langue	Total
Compréhension Conte d'Andersen : la petite fille aux allumettes	22 pts (12 questions)	18 pts (8 questions)	2 pts (2 questions)	42 pts
Grammaire	–	–	23 pts (4 questions)	23 pts
Production écrite (suite du conte)	5 pts	–	8 pts	13 pts
Total	27 pts	18 pts	33 pts	78 pts

De manière générale, les formats des questions sont variés : QCM, questions à réponse courte, questions ouvertes.

L'épreuve comporte également une unité thématique : ici le conte d'Andersen (la petite fille aux allumettes).

Si l'on regarde de plus près les différentes questions (selon notre interprétation), elles peuvent être réparties de la manière suivante.

Tableau 4.11. Composition à l'intérieur des domaines

	Composantes	Objectifs	Nombre de questions	Nombre de points
Compréhension de l'écrit	*Contexte*	- Connaître les différentes parties d'un livre (couverture, titre, etc.)	1	3
	Contenu	- Dégager le thème principal du texte - Repérer dans un texte l'élément d'information recherché - Comprendre une information implicite (interpréter)	13	16
	Planification	- Retrouver l'ordre des principales phases du conte	1	2
	Textualisation	- Comprendre le sens d'un mot, d'une phrase - Retrouver un verbe (justifier un élément) ou un autre type de mot - Interpréter les reprises pronominales et nominales - Identifier le rôle d'un élément de ponctuation - Utiliser la ponctuation - Identifier un type de discours (p. ex. monologue)	12	19
Production écrite	*Contexte*	–	–	–
	Contenu	- Raconter qq chose qui ne figure pas dans le texte	1	2
	Planification	- Être cohérent (respecter l'ordre chronologique, établir des relations de causes)	1	3
	Textualisation	- Respecter les contraintes orthographiques et syntaxiques (et temps des verbes)	2	8
Grammaire	*Grammaire*	- Connaître les catégories grammaticales - Utiliser les anaphores (pronoms ou groupes nominaux)	2	14
	Orthographe (pas dans cette partie mais dans Compréhension)	- Copier une phrase correctement - Composer une phrase correcte	2	2
	Conjugaison	- Transposer les temps (passé simple → passé composé) - Conjuguer au temps demandé	2	9
	Vocabulaire	–		

Sous le terme de grammaire, on retrouve le domaine structuration utilisé à l'école primaire. Il n'y a pas dans cette partie de questions portant à proprement parler sur le vocabulaire. Pour ce sous-domaine, les questions s'y rapportant sont davantage intégrées à la compréhension (compréhension de vocabulaire par le contexte). Pour ce qui relève de l'orthographe, il n'y a pas non plus de questions ou de points d'orthographe mais il est pris en compte en production écrite et en compréhension quand on demande aux élèves de recopier une phrase ou de formuler une réponse sous forme de phrase.

Comparaison des deux épreuves

Les deux épreuves analysées présentent un certain nombre de points communs : une unité thématique pour l'ensemble du français, un certain nombre de domaines abordés (compréhension de l'écrit, production écrite, structuration ou grammaire), des questions de format variés (QCM, fermées à réponse courte, ouverte, etc.).

Les trois principales différences entre les deux épreuves concernent les points suivants :

- un temps différent : le temps prévu pour l'épreuve est beaucoup plus long au primaire (5h15 en tout) *vs* 95 mn. Précisons qu'au primaire, une mise en contexte, notamment pour la production écrite, est prévue. Par ailleurs, en 6P, on peut considérer qu'il y a deux épreuves de français (l'une en français I et l'autre en français II) alors qu'en 7e il n'y en a qu'une, le français technique étant intégré à la compréhension et la production écrite ;
- si la partie compréhension d'un texte écrit est assez semblable du point de vue du nombre et du type de questions (on trouve un peu plus de questions d'interprétation dans l'épreuve de 7e mais pas de question pour identifier le but du texte), il y a beaucoup plus de différences concernant la production écrite (grille d'évaluation détaillée au primaire, moins au CO) et surtout la structuration ou grammaire ;
- enfin, ce qui relève de la table de spécification (très détaillée au primaire, par objectif, alors qu'au CO elle est davantage implicite et se base surtout sur les trois composantes *Contenu*, *Moyens langagiers* et *Langue*) et la manière d'élaborer le barème sont très différents. Au primaire, en lien avec la table de spécification (appuyée par les prétests), le barème est conçu a priori tandis qu'au CO, il est établi a posteriori.

Évaluation des mathématiques en 6P et 7e

Quelques comparaisons de forme

La comparaison des épreuves de 6P et de 7e CO met d'emblée en évidence une différence importante en ce qui concerne le nombre de questions et le temps imparti.

Tableau 4.12. Comparaison d'épreuves de mathématiques de 6P et de 7e CO

	6P	7e CO A-H	7e CO B-C
nombre de questions	13	18	18
nombre d'items	46	55	54
nombre total de points	38	72	67
temps de passation	2 fois 120 minutes	95 minutes	95 minutes

La plupart des contenus mathématiques de 7e CO sont ceux de 5P-6P mais sont approfondis et généralisés. Quelques autres, approchés à l'école primaire, font l'objet d'une construction et d'une consolidation. Les nouveaux savoirs concernent essentiellement les fractions ordinaires et le calcul littéral.

Les contenus mathématiques en 6P et le champ couvert par l'épreuve

Le plan d'études romand pour les degrés 1 à 6 comporte six domaines :

- formes géométriques,
- repérage dans le plan et dans l'espace,
- transformations géométriques,
- nombres entiers naturels,
- nombres réels et mesures,
- opérations, fonctions et linéarité.

L'épreuve 6P 2008 répartit les trois premiers domaines dans la catégorie « espace » et les trois suivants dans celle du « nombre ». Elle est composée comme suit :

- La catégorie « espace » est évaluée par des questions relatives
  - aux transformations géométriques (symétrie axiale),
  - aux figures géométriques,
  - au repérage dans le plan (coordonnées).
- La catégorie « nombre » est évaluée par des questions relatives
  - à la droite numérique,
  - aux fonctions et à leurs représentations,
  - à la connaissance des nombres entiers naturels avec compositions et décompositions additives et multiplicatives, opérations, suites numériques,
  - aux problèmes additifs et multiplicatifs (addition, soustraction, multiplication, division),
  - à la notion de proportion,
  - à la mesure (notion et calcul d'espace, de périmètre, d'aire et de volume) (5 questions).

Les contenus mathématiques en 7e CO et le champ couvert par l'épreuve

Les moyens d'enseignement romands en mathématiques du CO comporte cinq domaines :

- géométrie,
- fonctions, logique et raisonnement,
- nombres et opérations,
- calcul littéral,
- grandeurs et mesures, analyse de données.

La table de spécification du CO (élaborée en 2008 et en vigueur à la rentrée 2008-2009) répartit ces domaines comme suit :

- nombre et opérations,
- proportionnalité,
- algèbre,
- grandeurs et mesures,
- géométrie,
- fonctions.

Ces différents domaines doivent être testés dans toutes les épreuves mais leur importance respective varie au cours des années de manière préétablie. En 9e seulement, l'initiation aux fonctions et le thème « calculatrice » sont testés.

En 2008, l'algèbre (calcul littéral) n'est pas testée en 7e.

L'épreuve de 7e comporte deux séries similaires où seuls les nombres varient (les questions sont les mêmes) pour éviter le copiage. Elle présente également deux formes différentes suivant qu'elle s'adresse aux regroupements A-H ou B-C. Celles-ci comprennent le même nombre de questions et un seul item de moins. En revanche, malgré l'apparente ressemblance des questions entre les deux regroupements, seules quatre d'entre elles sur 18 sont exactement les mêmes. Toutes les autres comportent des différences qui consistent en données et écritures numériques plus simples (p. ex., un code fractionnaire est remplacé par un code décimal et vice-versa suivant la difficulté), les opérations arithmétiques sont simplifiées (p. ex. une division de deux nombres décimaux est remplacée, en B-C,

par une division de deux entiers). La mesure, la géométrie et la proportionnalité présentent aussi des problèmes simplifiés.

L'épreuve 2008 est composée comme suit :

- Les nombres et les opérations sont évalués par des questions relatives
  - à l'ensemble N (entiers positifs) : numération, opérations, propriétés des opérations, suites numériques ;
  - à l'ensemble Z (entiers relatifs) : numération, opérations ;
  - à l'ensemble R (réels) : numération, opérations.
- La proportionnalité est évaluée par des questions relatives
  - à la compréhension de la notion ;
  - aux fractions ordinaires (celles-ci étant également testées dans les questions relatives à la numération (R).
- La géométrie est évaluée par des questions relatives
  - au repérage dans le plan (coordonnées, équidistance, …) ;
  - aux formes géométriques et à leurs propriétés respectives (notions d'angle, de médiatrice, de bissectrice, …).
- La mesure est évaluée par des questions relatives
  - à la mesure du temps ;
  - à la mesure de figures (périmètre, aire) ;
  - aux transformations d'unités.

En 2008, l'algèbre (calcul littéral) n'est pas évaluée en 7^{e} ; elle constitue une nouvelle discipline par rapport à la 6P et n'a peut-être pas été abordée par toutes les classes au moment de l'épreuve. En revanche, l'absence de problèmes relatifs aux fonctions crée une rupture par rapport à l'école primaire.

Comparaison entre le contenu des épreuves de 6P et de 7^{e} CO

En 6P, chaque question appartient explicitement à l'une de deux catégories (espace et nombre). Les points obtenus à l'épreuve sont répartis selon ces deux catégories, de même que les résultats obtenus. Il n'y a pas de traitement statistique par question. Au CO, le nombre de points de chaque question d'EVACOM Mathématiques est relevé et saisi mais n'est pas attribué explicitement à un domaine sur le document « critères de correction » fourni aux enseignants. C'est l'équipe rédactionnelle des épreuves qui dispose d'un cadre de travail, lequel comporte une mention « pondération » qui répartit les points d'une épreuve entre les différents domaines et leur importance respective dans le programme de chaque degré. Quel que soit le degré, les rédacteurs d'épreuve doivent veiller à ce qu'un tournus règle, au fil des années, l'évaluation de thèmes moins souvent abordés ou évalués.

Dans le but de comparer les épreuves de 6P et de 7^{e} CO, nous avons soumis celles de 7^{e} à la même catégorisation que celle de 6P (nombre ou espace).

Tableau 4.13. Répartition des items en fonction de la catégorisation utilisée au primaire dans les épreuves de 6P et de 7e

	nombre	espace
6P	42 items en 10 questions 30 points	4 items en 3 questions 8 points
7e CO A-H	45 items en 15 questions 56 points	10 items en 4 questions 16 points
7e CO B-C	44 items en 15 questions 55 points	10 items en 4 questions 12 points

N.B. Quelques questions comportent des items relevant de plusieurs domaines.

La proportion respective de points attribués à la résolution des questions appartenant aux deux catégories « nombre » et « espace » est sensiblement la même en 6P et en 7e (un peu plus d'un quart des points attribués à l'espace ; un peu moins en regroupement B-C).

Nous avons également soumis l'épreuve de 6P à la répartition des questions selon les domaines évalués en 7e.

Tableau 4.14. Répartition des items en fonction de la catégorisation utilisée au CO dans les épreuves de 6P et de 7e

	nombre, opérations	proportionnalité	mesure	géométrie
6P	5 questions 28 items 14 points	3 questions 8 items 10 points	3 questions 6 items 8 points	3 questions 4 items 6 points
7e CO A-H	7 questions 24 items 29 points	5 questions 10 items 12 points	4 questions 11 items 15 points	4 questions 10 items 16 points
7e CO B-C	7 questions 24 items 29 points	5 questions 10 items 12 points	4 questions 10 items 14 points	4 questions 10 items 12 points

N.B. Quelques questions comportent des items relevant de plusieurs domaines.

C'est en numération que la proportion des points est la plus considérable quels que soient l'année et le regroupement (environ 2/5). Cette proportion est judicieuse si l'on tient compte des nombreux champs notionnels inclus dans ce domaine.

La proportionnalité est diversement évaluée. La 6P évalue la notion de proportion mais explore aussi des situations relatives aux « fonctions », ce qui n'est pas le cas en 7e, en 2008.

De ce fait, le domaine proportionnalité, fonctions et linéarité de 6P atteint les 26% des points contre environ 17-18% en 7e.

La mesure qui porte essentiellement sur les notions et calculs de périmètres et d'aires quels que soient l'année et le regroupement est aussi similairement évaluée (environ 1/5 des points).

En géométrie, la proportion des points est la plus élevée en 7e A-H (22%) et relativement proche entre la 6P (16%) et la 7e B-C (18%).

Les différents domaines mathématiques sont tous évalués, chacun selon la place qu'il tient dans le programme. Seul le champ relatif aux fonctions (linéaires par exemple) n'apparaît pas en 7e en 2008,

mais il faudrait poursuivre ce type de comparaison sur plusieurs années pour avoir une photographie plus équitable des contenus évalués, d'autant que, comme le montre le premier tableau comparatif, l'épreuve de 7e est déjà très longue par rapport au temps imparti.

Comparaison entre les types de questions

La table de spécification concernant l'épreuve de 6P 2008 comporte une autre dichotomisation : les questions sont réparties entre « problèmes d'application » et « situations problèmes ».

Les problèmes d'application consistent à recourir à un ou plusieurs algorithmes, formules, procédures routinières ou conventions élémentaires. Certains de ces problèmes nécessitent des connaissances affirmées et pointues.

Les situations-problèmes impliquent de comprendre et d'interpréter les données, de tirer des conclusions, d'inférer des procédures plus ou moins complexes, d'expliciter ou de justifier sa réponse. A un niveau plus élevé, il peut s'agir d'établir une stratégie pour gérer une situation inconnue.

Dans le but de comparer les épreuves, nous avons réparti les questions de l'épreuve de 7e CO selon ces deux types de problèmes.

Tableau 4.15. Répartition des items en fonction du type de question dans les épreuves de 6P et de 7e

	problèmes d'application	situations-problèmes
6P	35 items en 5 questions 13 points	11 items en 9 questions 25 points
7e CO A-H	27 items en 7 questions 27 points	27 items en 12 questions 45 points
7e CO B-C	28 items en 8 questions 28 points	25 items en 11 questions 39 points

En 6P, la majorité des points est consacrée aux question de type situation-problème (65%).

La 7e CO attribue presque la même proportion de points aux situations-problèmes. Notons la tendance, récurrente dans l'enseignement, d'alléger les épreuves en situations-problèmes pour les élèves moins doués. Toutefois, l'épreuve 2008 présente un décalage assez faible : 63% des points attribués aux situations-problèmes dans le regroupement A-H contre 58% dans le regroupement B-C.

Correction et barèmes respectifs

Quel que soit le degré, les enseignants chargés de la correction reçoivent un document qui précise la façon d'attribuer les points pour chaque question et pour chaque item.

En 6P, le document remis aux enseignants indique le barème, c'est-à-dire le nombre de points obtenus est attribué une note. Il est établi en fonction des objectifs, a priori.

Au CO, le document stipule que les enseignants trouvent notes et barèmes sur intranet. Ceux-ci sont établis après la passation de l'épreuve, sur la base des résultats des élèves.

En résumé, on constate trois points de rupture entre l'épreuve de 6P et les épreuves de 7e CO :

- le contenu : le CO n'évalue pas les fonctions, ce qui crée une rupture dans la continuité du programme ;
- le nombre de questions par domaines : l'épreuve de 6P a plus de la moitié moins d'items (seulement 4 items pour tout le champ spatial) ;
- le temps imparti : une minute et demie par item en 7e contre plus de 5 mn en 6P ; même si certains d'entre eux peuvent être vite résolus, il reste peu de temps pour le raisonnement, surtout pour la population de niveaux B-C.

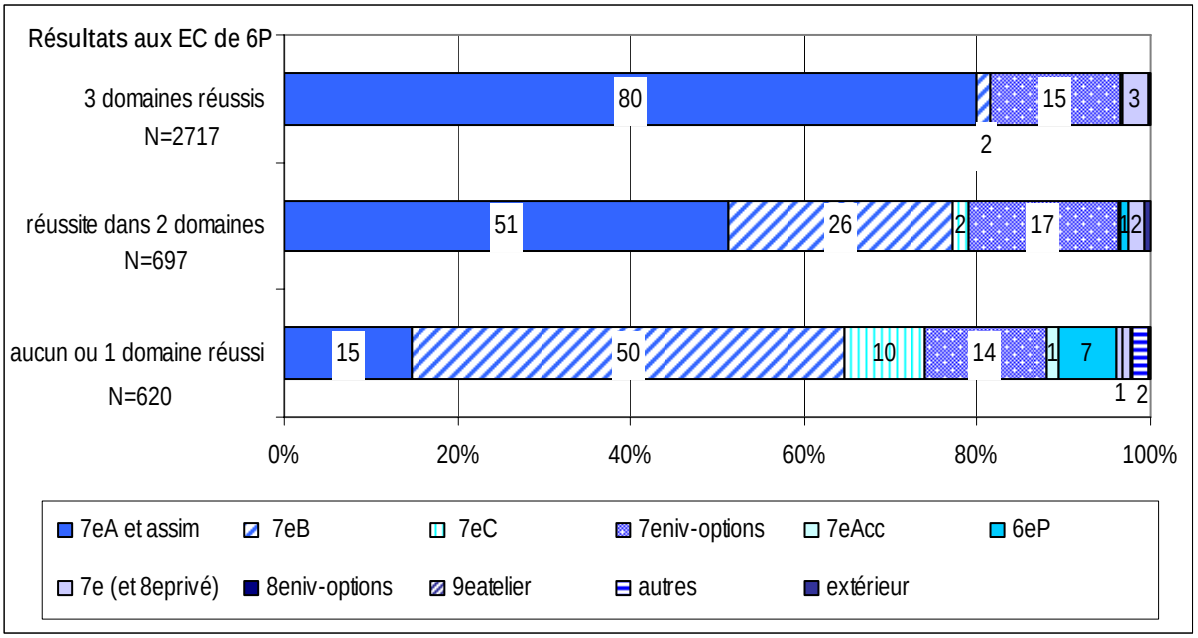
La grande différence réside dans le nombre de questions que comportent les épreuves respectives par rapport au temps attribué aux élèves pour les résoudre. Une question se pose dès lors : s'il est normal que la difficulté s'accroisse et que le temps imparti diminue, une seule année dans le développement cognitif et scolaire des élèves permet-elle un décalage aussi important entre les épreuves ? Sur le plan didactique, peu de temps pour réaliser une longue tâche signifie que l'on favorise l'accumulation des savoirs au détriment du raisonnement, ce qui risque de constituer un hiatus par rapport à la conception des mathématiques préconisée par le futur Plan d'études romand.

5. Liens entre les résultats aux épreuves cantonales de 6P et l'orientation au CO : par cours de la 7e à la 9e

Afin de voir concrètement les liens entre les résultats aux épreuves cantonales et l'orientation, ou en d'autres mots observer la prédictivité (ou fonction pronostique) des EC par rapport à l'orientation des élèves, nous avons essayé d'observer le parcours des élèves qui ont passé les épreuves cantonales en fin de 6P en 2006 et avons regardé leurs parcours les trois années suivantes, à titre d'exemple. Il faut préciser que les épreuves cantonales ne sont pas conçues au départ pour être prédictives. De plus, il va de soi que les résultats aux épreuves cantonales ne sont pas censés déterminer à eux seuls l'orientation des élèves mais il paraissait intéressant de mettre en relation ces résultats avec les orientations futures. Il serait utile de faire le même type d'analyse avec les évaluations de fin de 6P réalisées par les enseignants.

Nous illustrerons tout d'abord par la situation des élèves en 7e année selon qu'ils ont atteint le seuil de réussite fixé par l'institution dans les trois domaines nécessaires pour l'orientation (français I, français II ou structuration, mathématiques), dans deux domaines sur trois (le plus fréquemment, français I et mathématiques) ou encore dans un seul domaine, voire aucun : sur l'ensemble des élèves environ deux tiers réussissent les trois domaines, 17% deux domaines et 15% un seul voire aucun des trois domaines.

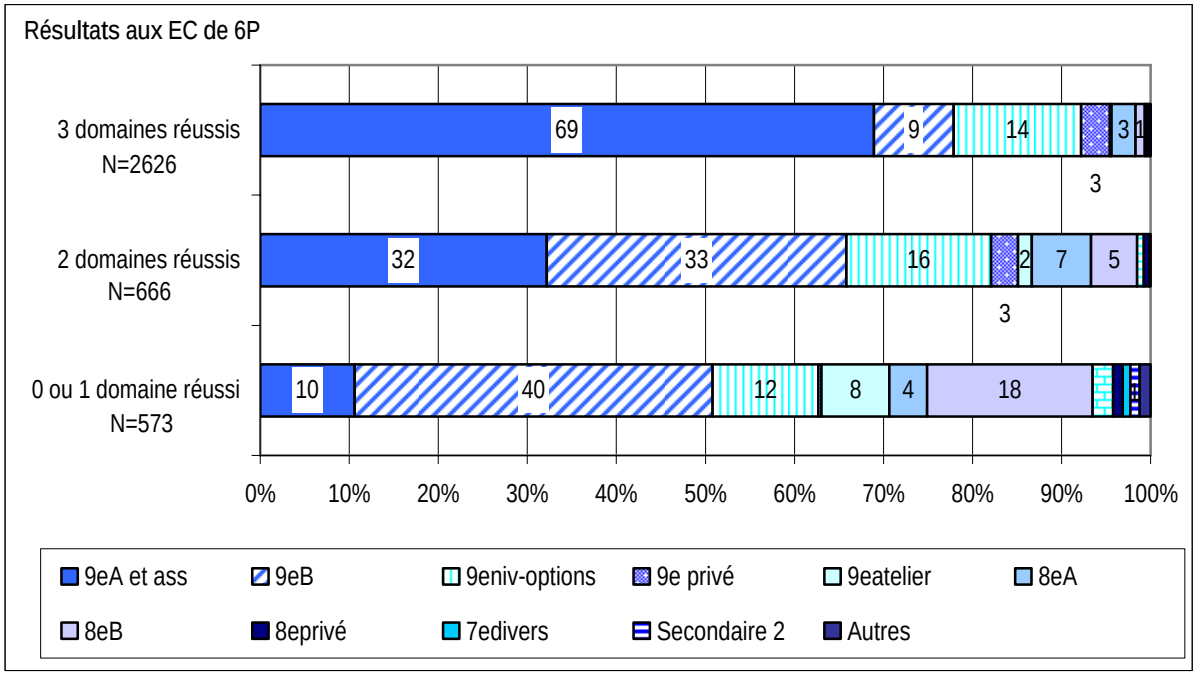
Graphique 5.1. Situation en 7e année (année 2006-2007)



Comme on peut le constater, on observe une certaine correspondance entre les résultats aux épreuves cantonales et l'orientation en 7e : 80% des élèves ayant réussi dans les trois domaines sont orientés en A ou assimilés (classes spéciales : sport et danse). Les 20% restants étant surtout composés d'élèves répartis dans les établissements à niveaux et options[40]. Si l'on observe les élèves n'ayant pas réussi les épreuves cantonales ou n'ayant réussi qu'un domaine, on constate que 50% d'entre eux se trouvent dans une 7e B, 10% dans une 7e C, 14% dans une classe hétérogène, 7% ont redoublé. Relevons la présence de 15% de ces élèves qui ont été malgré tout orientés en 7e A en fonction de leurs résultats de l'année. Une situation intéressante est celle des élèves qui avaient réussi deux domaines sur trois : un peu plus de la moitié d'entre eux ont été orientés en A et 28% en B ou en C.

Si l'on regarde la situation de ces élèves deux ans plus tard, on obtient la situation suivante :

Graphique 5.2 Situation en 9e année (année 2008-2009)



La situation deux ans plus tard montre une prédictivité[41] un peu moins bonne. Les élèves qui avaient réussi sont proportionnellement moins nombreux en A (69% *vs* 80% en 7e), 9% sont en 9e B, 14% dans des établissements à niveaux et options, le reste se répartissant entre le privé, un redoublement en 8e A ou B ou dans diverses orientations (certaines dans le secondaire II). Pour ce qui concerne la différence de pourcentage d'élèves en 7e A et en 9e A, il faut préciser que de manière générale, indépendamment des résultats aux épreuves cantonales de 6P, la proportion d'élèves de A est moins élevée en 9e qu'en 7e (par exemple les élèves entrés en 7e A en 2006-07 représentaient 78% environ des élèves scolarisés dans des établissements à regroupements différenciés et représentent en 2008-2009 69% environ en 9e A).

A l'autre extrême, chez les élèves qui n'avaient réussi qu'un seul domaine, voire aucun, 40% se trouvent en 9e B, 10 en 9e A, 12% dans des établissements à niveaux et options, 22% sont encore en

[40] Pour ces trois établissements, en 7e année les élèves se trouvent dans des classes hétérogènes. En 8e et 9e, ils ont des niveaux (en mathématiques et allemand) ainsi que des options. Malheureusement, nous n'avons pu récolter les informations sur les options choisies.

[41] Soulignons toutefois que ce n'est pas l'objectif des épreuves cantonales d'être prédictives mais il est intéressant de voir quel est le lien entre les résultats aux EC en fin de primaire et le parcours des élèves à la fin du CO (trois ans plus tard).

8e B voire A et 8% ont été orientés dans une classe atelier. Quant à ceux qui avaient réussi deux domaines sur trois, leur situation est contrastée puisqu'un tiers se trouve en 9e A et un autre tiers en 9e B.

On trouvera, dans l'annexe 5, les parcours de la 7e à la 9e pour les trois profils de résultats qui peuvent être résumés comme suit dans le tableau 5.1.

Comme on peut le constater, les parcours sont variés. Dans la catégorie *autres*, on trouve des parcours contrastés allant d'élèves sortis du système à d'autres passés dans d'autres types de structures (spécialisé, préapprentissage ou secondaire II comme le SCAI, des apprentissages ou encore dans de très rares cas, le Collège). Tout en restant prudent étant donné qu'il ne s'agit que d'une volée et que nous n'avons pas pu récolter les orientations des élèves se trouvant dans les cycles à niveaux et options, on peut souligner que globalement l'orientation s'est plutôt bien faite puisqu'un certain nombre d'élèves ayant réussi ou au contraire n'ayant pas réussi se retrouvent où on pouvait les attendre compte tenu de leurs résultats aux épreuves en fin de 6P. Il est intéressant de constater la grande hétérogénéité des parcours des élèves qui avaient réussi deux domaines sur trois.

Les redoublements sont plus fréquents chez les élèves qui n'avaient réussi que dans un seul domaine voire dans aucun. Toutefois, dans de rares cas, on peut observer une réorientation positive proportionnellement plus fréquente dans cette catégorie d'élèves.

Globalement, les épreuves cantonales semblent avoir une certaine prédictivité par rapport à l'orientation au CO même au bout de trois ans.

Tableau 5.1. Parcours des élèves ayant passé les épreuves cantonales 2005-06 de la 7e à la 9e année

	0-1 domaine réussi		2 domaines réussis		3 domaines réussis	
	Nombre	*%*	*Nombre*	*%*	*Nombre*	*%*
Parcours linéaires						
7e A / 8e A / 9e A	41	6.6	189	27.1	1783	65.6
7e B / 8e B / 9e B	161	26.0	105	15.0	22	0.8
7e C / 8e B / 9e B	24	3.9				
Parcours sans indication d'orientation mais sans redoublement						
7e privé / 8e privé / 9e privé	1	0.2	11	1.6	72	2.6
7e niveaux-options / 8e n-o / 9e n-o	61	9.8	104	14.9	370	13.6
7e A / 8e A / 9e niveaux-options					4	0.1
Parcours avec réorientations « sélections »						
7e A / 8e A / 9e B	13	2.1	47	6.7	111	4.1
7e A / 8e B / 9e B	14	2.3	56	8.0	97	3.6
Parcours avec réorientations « promotions »						
7e B / 8e A / 9e A	15	2.4	17	2.4	4	0.1
7e B / 8e A / 9e B	9	1.5	12	1.7	2	0.1
Parcours avec redoublements						
7e A / 8e A / 8e A	6	1.0	22	3.2	50	1.8
7e A / 8e B / 8e B	2	0.3	4	0.6	6	0.2
7e A / 7e A / 8e A	5	0.8	14	2.0	18	0.7
7e A / 7e A / 8e B	1	0.1			4	0.1
7e B / 8e A / 8e B					1	0.0
7e B / 8e A / 8e A	6	1.0	3	0.4		
7e B / 7e B / 8e B	28	4.5	7	1.0	4	0.1
7e B / 8e B / 8e B	1	0.2	1	0.1		
7e B ou C / 8e B / 9e atelier	25	4.1	6	0.8		
7e C / 8e B / 8e B	11	1.8	1	0.1		
7e niveaux-options / 8e n-o / 8e n-o	4	0.6	3	0.4	4	0.1
6P / 7e B / 8e B	12	1.9	2	0.3		
Autres	180	29.0	93	13.3	165	6.1
Total	620	100%	697	100%	2717	100%

6. Une évaluation à part : la préorientation

Lors de la discussion du mandat, la question de la préorientation a été évoquée dans la mesure où il s'agit d'une forme d'évaluation externe mise en place par l'institution. Par préorientation, il faut entendre une opération dépendant du CO pour le passage en 7e. Elle s'adresse aux élèves de 6P scolarisés à Genève dans les écoles publiques (y compris ceux qui viennent de l'enseignement spécialisé) mais également à des élèves scolarisés dans les écoles privées ainsi que ceux qui viennent d'autres cantons ou d'autres pays. Cette procédure s'applique également en 8e et 9e pour les élèves provenant d'écoles privées ou d'autres cantons ou encore de l'étranger. La préorientation représente en tout 4500 élèves par an.

L'admission au CO des élèves issus des écoles primaires genevoises (2008) éditée par le Service de la scolarité de la Direction générale du CO rappelle que l'admission au CO se fait sur la base des résultats de 6P. Il est précisé que « l'évaluation de fin de 6P est du ressort de l'école primaire alors que l'admission au CO et l'orientation des élèves sont de la responsabilité des directions et des établissement du CO » (2008, p. 7). D'autres éléments que les résultats de l'année peuvent être également pris en compte en cas de situations plus délicates : le rapport de l'enseignant primaire, l'avis de l'élève et de ses parents, les indications fournies par les tests de raisonnement[42] ainsi que les résultats aux épreuves cantonales. Concernant les tests de raisonnement, ils sont administrés chaque année à l'ensemble des élèves de 6P par des psychologues. Ils se composent de quatre tests : logique générale, spatial, verbal et numérique.

Nous avons interrogé deux personnes à ce sujet : le coordinateur des épreuves cantonales du primaire, adjoint à la direction et le responsable de la préorientation au CO.

Le premier interlocuteur regrette que la préorientation soit entièrement prise en charge par le CO, contrairement à la situation existant dans d'autres cantons (p. ex. dans le canton de Fribourg). Par ailleurs, il souligne que les tests de raisonnement en particulier sont peu pris en compte, utilisés surtout dans des cas un peu limites (dérogation pour entrer dans le regroupement A, passage de l'enseignement spécialisé au CO ou au contraire dans les EFP, notamment). Étant donné le temps nécessaire pour leur passation et l'utilisation des résultats, il estime que ces tests ne devraient être passés qu'aux élèves pour lesquels on se pose des questions.

Le responsable de la préorientation a une vision un peu différente. Il rappelle que la préorientation consiste à recueillir des informations sur tous les élèves qui vont arriver au CO, informations pédagogiques, psychologiques et rapports du maitre. Il s'agit non seulement du passage de la 6P au CO mais également de l'orientation des élèves des classes spécialisées : ceux qui ont un fort potentiel mis en évidence dans les tests de raisonnement entreront au CO, tandis que ceux qui ont un potentiel relativement bas seront dirigés vers les EFP. Si ces tests sont utilisés pour assez peu d'élèves, ils pourraient l'être davantage. Par ailleurs, il est difficile de savoir avant la fin de l'année pour quels élèves ces tests seraient nécessaires. De manière plus générale, on pourrait imaginer que les psychologues reçoivent les orientations de tous les élèves de 6P pour la 7e et les résultats aux tests de raisonnement et qu'ils confrontent les deux types d'information : par exemple, est-ce que pour un élève orienté vers le regroupement A, cela correspond à son potentiel ? Une recherche est actuellement en cours en collaboration avec la FPSE (Th. Lecerf) avec le responsable de la préorientation concernant la validité prédictive des tests et les résultats scolaires. Cette étude apportera un éclairage intéressant sur les liens entre les tests de raisonnement et les résultats scolaires.

7. L'avenir et la place des différentes évaluations externes : tests HarmoS, épreuves de référence, évaluations cantonales et communes et PISA

Une partie du mandat consistait également à s'interroger sur la cohérence des différentes évaluations externes (cf. description détaillée des différentes évaluations externes existantes dans l'annexe 6) et notamment les futurs tests HarmoS ainsi que les épreuves de référence romandes. Il est pour l'instant difficile de se prononcer clairement étant donné que ces deux types d'évaluation n'existent pas encore. Rappelons tout d'abord ce qui est prévu avant de donner la parole aux différents acteurs.

Tout d'abord, les tests HarmoS se situent au plan national dans le cadre de l'harmonisation de la scolarité. Des standards de formation ont été proposés dans quatre domaines : la langue d'enseignement (L1), les langues secondes (L2 et L3), les mathématiques et les sciences, à trois moments-clés du cursus scolaire, à savoir en 4e (anciennement 2e année), 8e (ex-6e) et 11e (ex-9e année,

[42] Ces tests de raisonnement sont considérés comme très utiles par l'actuelle directrice de l'enseignement du CO. Ils permettent de déterminer le potentiel des élèves et complètent les résultats de l'année (évaluations des enseignants et épreuves cantonales).

c'est-à-dire à la fin de la scolarité obligatoire) et devraient faire l'objet d'une consultation. Ces standards de formation se basent sur des compétences minimales que les élèves devraient avoir acquises dans ces quatre domaines. Il est prévu par la CDIP que des tests de référence soient développés dans le cadre d'un monitorage national pour vérifier l'atteinte de ces standards. L'évaluation qui a pour but d'évaluer le fonctionnement du système se ferait sur un échantillon d'élèves. Elle devrait avoir lieu à partir de 2011.

Pour ce qui concerne les épreuves de référence romandes, la CIIP prévoyait au départ la création d'épreuves de type bilan à la fin de chaque cycle. Elles pourront être réalisées à l'aide d'une banque de données romandes de tâches. Elles devraient permettre de vérifier l'atteinte des objectifs du plan d'études par tous les élèves. Elles sont directement liées au Plan d'études romand (PER) actuellement en consultation. Le projet, sous la responsabilité de l'IRDP, a démarré à l'automne après des travaux préalables et devrait s'achever en 2010. Si l'intention d'élaborer des épreuves de références romandes communes reste d'actualité, leur objectif n'a pas encore été déterminé, étant donné les attentes diverses des différents cantons (évaluation certificative, diagnostique, bilan ou encore au service de l'élève, de l'enseignant ou du système), et fait encore l'objet de discussions au sein de la CIIP.

Compte tenu de l'avancement de ces projets et des informations reçues, les différentes personnes interrogées ont souvent eu de la peine à nous répondre.

Enfin, PISA est une enquête internationale cyclique de l'OCDE qui vise, tous les trois ans, l'évaluation des compétences des jeunes de 15 ans en lecture, mathématiques et sciences. Lors de chaque enquête un des trois domaines fait l'objet d'une étude plus approfondie. Depuis le début de cette enquête, en 2000, la Suisse a élargi l'échantillon des élèves afin de bénéficier de données permettant leur comparaison à la fin de la scolarité obligatoire (9^{e} année) au niveau des régions linguistiques et des cantons qui le désiraient. L'ensemble des cantons romands, dont Genève, ont saisi cette occasion et évaluent les compétences des élèves de 9^{e}. Concrètement tous les trois ans, environ 1500 élèves genevois sont interrogés à l'aide des instruments PISA (deux heures de test et un questionnaire aux élèves de 45 minutes ainsi qu'un questionnaire destiné aux écoles). Cette enquête permet ainsi de cerner les compétences des élèves dans les trois domaines testés ainsi que l'impact des caractéristiques individuelles, scolaires et de contexte sur les performances des élèves. Signalons que cette évaluation ne permet pas de donner des résultats individuels des élèves. Il ne s'agit donc pas d'une évaluation visant à mesurer les acquis en fonction d'un programme scolaire ou à les certifier mais d'une évaluation système visant plutôt le pilotage des systèmes.

Au niveau du secrétariat général où les informations sont les plus précises (compte tenu du fait que le secrétaire général est justement responsable de cette question), il ressort qu'il est important que les épreuves genevoises soient en phase avec les deux autres types d'évaluation nationale et romande. Les tests HarmoS semblent avoir une fonction relativement claire : ils se situent au niveau du pilotage du système. Pour les épreuves de référence, les choses sont plus délicates. Il s'agit d'éviter les doublons, de ne pas créer une « armada » d'épreuves cantonales, romandes, etc. et d'utiliser au mieux l'expertise développée à Genève en matière d'élaboration d'épreuves. On pourrait imaginer de planifier par exemple une épreuve romande en mathématiques en 8^{e} primaire, ce qui donnerait des éléments intéressants en termes d'adéquation par rapport au programme. Les épreuves genevoises sont une espèce d'épreuve de référence. La division pourrait peut-être se faire au niveau de la fonction, tester les élèves ou le système.

A *l'école primaire*, pour la directrice de l'enseignement (comme pour la plupart des personnes interrogées), les tests HarmoS se situent au niveau de l'évaluation du système : par contre pour ce qui concerne les épreuves de référence romandes, cela dépend beaucoup de leur but : si elles peuvent répondre à cette attente d'évaluation des élèves et qu'elles sont bien faites, elles pourront remplacer les épreuves cantonales ou communes. La difficulté sera plus ou moins grande selon les domaines : par exemple, en allemand, la focalisation n'est pas la même à Genève et dans d'autres cantons. Toutefois, il n'y aurait pas de raison qu'on ne puisse pas le faire si le Plan d'études romand est introduit.

Par contre, les épreuves cantonales ou communes sont des outils d'évaluation utiles qui donnent un regard sur le niveau d'acquisition des élèves, regard externe mais interne à l'institution. Elles sont également utiles car elles ont un effet de régulation notamment pour les jeunes enseignants.

Le coordinateur des épreuves cantonales se réjouit de l'existence des tests HarmoS, en tant qu'épreuves-système dont la Suisse et les cantons manquaient jusque là. Il est également content que l'on envisage un espace romand d'enseignement avec des épreuves standardisées pour toute la Suisse romande qui pourraient remplacer les épreuves cantonales. Toutefois, cela risque d'être difficile étant donné les différences entre cantons romands du point de vue des législations et surtout des pratiques. Chez les concepteurs, le flou est assez présent, surtout en ce qui concerne les épreuves romandes. La question se pose notamment par rapport à la marge de liberté cantonale en lien avec le PER. En allemand, en particulier, on s'interroge étant donné que dans d'autres cantons, l'enseignement est davantage axé sur le vocabulaire. Du côté du SEDEV, on estime que les épreuves de référence pourraient être assez proches des épreuves cantonales et les remplacer.

Au *cycle d'orientation*, la directrice de l'enseignement souligne la multitude d'épreuves auxquelles il faut ajouter l'enquête internationale PISA. Toutes ces épreuves et tests constituent un investissement important pour les élèves. Elle se demande comment utiliser au mieux et concilier ces différentes sources d'information (dont les résultats peuvent s'avérer contradictoires). Les tests HarmoS sont identifiés comme servant au pilotage et au monitorage du système, sont centralisés et réguleront le système au niveau national. Quant aux épreuves de référence romandes, elle rappelle qu'au début elles étaient censées être de type diagnostique. Pour elle, il devrait y avoir soit des épreuves de référence romandes soit des épreuves communes ou cantonales. Sinon il y aura des problèmes de ressources. Par ailleurs, l'introduction du PER justifierait ce choix. On pourrait éventuellement avoir des épreuves de référence dans certaines disciplines : français, anglais, allemand, mathématiques, par exemple et avoir à côté des épreuves cantonales dans d'autres telles que le latin, l'histoire, la géographie, etc.

On soulignera un avis un peu différent de celui des autres interlocuteurs, celui du responsable du SEC qui s'interrogeait sur l'accueil d'épreuves de référence romandes, même si des concepteurs genevois sont délégués pour participer à leur élaboration.

Pour le responsable de la préorientation, si les épreuves de référence romandes ont les mêmes finalités, on pourrait imaginer que les épreuves cantonales porteraient plutôt sur des champs plus difficiles à évaluer avec des épreuves papier-crayon, telles que les compétences de résolution de problème, l'expression orale, les expérimentations, etc.

Pour les commissaires de français, une des questions qui se pose est le découpage annuel, le PER n'étant pas découpé par année mais par cycle. En allemand, on regrette que les épreuves soient construites par des experts externes et non des enseignants avec une pratique de classe. Par ailleurs, elles estiment avoir de l'avance dans la réflexion car la présidente de groupe a participé au PER et le système utilisé dans le cadre de leur enseignement est le cadre européen des langues. La coordination sera plus facile.

En mathématiques, les commissaires proposent qu'il y ait une partie commune et une autre spécifique.

Enfin, en physique, on relève que les épreuves communes n'ont pas le même rôle que les épreuves de référence : les premières servent à la certification des élèves tandis que les secondes ont pour but de voir si l'on remplit les objectifs du PER.

Synthèse et réflexions à propos de la nouvelle organisation de l'évaluation cantonale et commune dans l'enseignement obligatoire

Dans cette synthèse, nous rappellerons d'abord brièvement les principaux résultats qui ressortent de cette étude. Ensuite, nous apporterons des éléments de réflexion en vue de la définition d'une nouvelle organisation et nous mettrons en évidence quelques contraintes et limites pour la mise en œuvre d'un tel dispositif.

Rappelons que le but de ce travail était de faire un inventaire des dispositifs d'évaluation externes en termes de points forts et de points faibles, de les mettre en perspective avec les nouveaux dispositifs d'évaluation prévus au niveau national et régional, de proposer une organisation transversale d'évaluation (et d'établir un plan de mise en œuvre).

L'analyse des documents disponibles montre que les épreuves cantonales et communes genevoises sont des épreuves qui sont administrées à l'ensemble des élèves du degré et de la discipline concernée. Elles visent à vérifier si les élèves atteignent les objectifs du plan d'études et à réguler l'enseignement. De plus, ces épreuves entrent dans la certification des élèves. Si ce type d'épreuves existe dans la plupart des cantons romands, notons toutefois que c'est à Genève que le nombre d'épreuves administrées est le plus élevé.

Pour réaliser cet inventaire différents groupes d'acteurs ont été interrogés. Les responsables du Département (le Secrétariat général), les directions des ordres d'enseignement (directrices de l'enseignement et responsables de l'évaluation cantonale ou commune), les concepteurs des épreuves ainsi que les enseignants. Par ailleurs, un échantillon d'épreuves (français et mathématiques) ont été analysées du point de vue docimologique.

Le dispositif des évaluations externes genevoises peut être caractérisé par un certain nombre d'éléments. Tout d'abord, comme nous l'avons relevé à plusieurs reprises, l'organisation est différente dans les deux ordres d'enseignement : à l'école primaire, les épreuves sont élaborées par des didacticiens-formateurs en collaboration avec des formateurs en évaluation et sous la responsabilité d'un coordinateur, expert en évaluation. Au CO, ce sont des enseignants avec une pratique de classe, qui sont dégrevés pendant un temps hebdomadaire, sous la responsabilité des présidents de groupe de la discipline et du responsable du SEC. Dans les deux cas, les personnes connaissent le programme mais dans un cas, ils ont développé une réflexion, des moyens d'enseignement et sont formateurs, tandis que dans l'autre cas, ils possèdent une connaissance quotidienne du terrain. Les compétences de ces deux types d'acteurs sont donc complémentaires.

L'ensemble des données ont été analysées en fonction de trois critères : la *qualité technique*, l'*utilité* des épreuves et leur *efficacité/efficience* : les points forts et les points faibles ainsi que les propositions d'amélioration seront mis en évidence pour chacun d'entre eux.

Qualité technique des épreuves

Les points forts

Si l'on regarde les points forts mis en évidence par les différents acteurs interrogés, certains sont communs aux deux niveaux d'enseignement. De manière générale, les différents acteurs interrogés (autorités, directions, concepteurs et enseignants) relèvent une bonne qualité des épreuves. L'analyse docimologique de quelques épreuves de mathématiques et français met en évidence une bonne fiabilité

des épreuves pour la certification des élèves et la régulation de l'enseignement ainsi qu'une bonne validité prédictive de l'épreuve de mathématiques (de 6P à 7e). Nous avons vérifié l'absence d'un biais dans l'épreuve de mathématiques qui aurait pu affecter sa validité de contenu (les élèves faibles en compréhension écrite ne sont pas significativement désavantagés pour la résolution de problèmes de mathématiques à longs énoncés).

Les éléments qui suivent se rapportent plus souvent à un ordre d'enseignement en particulier. Ainsi, *au primaire*, les différents acteurs soulignent l'existence de prétests, de tables de spécification qui permettent notamment d'identifier les objectifs évalués. Certains enseignants relèvent également le caractère objectif de ces épreuves, comparativement à celles élaborées par les enseignants pour leurs élèves.

Ces épreuves sont perçues comme étant un bon reflet de ce qui est attendu.

Les consignes de passation et de correction sont également jugées comme bien expliquées.

Enfin, l'expérience (et la stabilité) des concepteurs d'épreuves est également considérée comme un point positif. Il faut souligner qu'ils ont reçu une formation spécifique à l'évaluation (donnée par L. Allal) et que le dispositif comprend à la fois des formateurs/didacticiens des disciplines concernées et des formateurs en évaluation (SEDEV), l'ensemble étant encadré par le responsable des épreuves cantonales, spécialiste en évaluation.

Au *cycle d'orientation*, d'autres éléments sont mis en évidence et sont dépendants des disciplines considérées : une certaine stabilité de contenu et du niveau de difficulté (cette dernière a été confirmée par l'analyse docimologique des épreuves en mathématiques), de bonnes conditions de passation et de correction, une certaine clarté de l'épreuve et des consignes de correction (notamment en anglais), une bonne couverture du programme (par exemple en anglais et en physique) on encore un bon équilibre à l'intérieur des épreuves pour la plupart des disciplines.

Les points faibles

Certains points faibles observés dans le dispositif actuel sont communs aux deux ordres d'enseignement : des différences au niveau des approches selon les disciplines (et la nécessité d'harmoniser ces approches), une fiabilité insuffisante pour évaluer des connaissances ou compétences détaillées[43] (par exemple en 2P ou dans une épreuve de français au CO). L'analyse docimologique a permis de montrer la nécessité d'avoir un certain nombre d'items pour obtenir une bonne fiabilité. Or, plus on cherchera à évaluer des connaissances/compétences détaillées (p. ex. dans le domaine de l'espace ou des nombres), plus on s'exposera au problème de fiabilité insuffisante (surtout pour situer un élève sur l'échelle du test). Les responsables des épreuves font face à un difficile compromis entre le nombre de questions nécessaires pour évaluer fidèlement les acquis des élèves et la contrainte de ne pas poser trop de questions (surtout pour les élèves les plus jeunes). De manière générale, l'analyse des résultats voire le recours aux prétests est un point faible aux deux niveaux. Au primaire, les prétests sont plus systématiques mais ils servent seulement à tester la formulation des questions et leur compréhension, aident à déterminer la table de spécification et permettent d'éliminer certains items trop faciles ou trop difficiles. En revanche, les prétests ne permettent pas pour l'instant d'évaluer la qualité technique des épreuves (p. ex. validité, fiabilité).

Pour les autorités, on relève un certain manque de standardisation des épreuves. La nécessité d'équipes pluridisciplinaires (compétences au niveau de la connaissance du programme, de la construction d'épreuves et de l'analyse des épreuves) et de la cohérence entre épreuves du primaire et celles du CO (objectifs, passations, contenu, etc.) sont également pointées.

La plupart des points faibles concernent plutôt un ordre d'enseignement. A *l'école primaire***,** un certain nombre d'éléments sont relevés par l'ensemble des acteurs interrogés (autorités, directions de l'enseignement, les concepteurs ou encore enseignants) : l'existence de biais liés aux conditions de

[43] Cela implique que l'évaluation des connaissances/compétences détaillées ne doit être utilisée que de manière indicative.

passation et de correction (les enseignants des élèves se chargeant eux-mêmes de ces deux opérations), la couverture de l'épreuve en lien avec le programme (notamment dans le cas de l'allemand, la production orale), l'absence d'enseignants dans les commissions d'épreuves. Les enseignants relèvent également le manque de stabilité au niveau du contenu d'une année à l'autre, un niveau d'exigences peu adapté (notamment pour le français en 2P), la manière de tester la production écrite jugée peu adéquate.

Au *cycle d'orientation*, on déplore souvent le manque de prétest selon les disciplines, faute de temps[44]. Il faut souligner qu'une étude réalisée dans les années 80 (Davaud, Hexel, Bain, 1983) évoquait déjà la nécessité de prétests. Les éléments suivants sont également mentionnés par les différents acteurs : l'absence de tables de spécification dans les épreuves de certaines disciplines[45], les conditions de correction (assurée par les enseignants des élèves considérés), pouvant donner lieu à des biais, le manque de stabilité et la formation des concepteurs estimée parfois insuffisante[46] (notamment lorsqu'on introduit l'évaluation de nouveaux objets), la façon dont le programme est pris en compte et la manière de tester (micro-objectifs évalués dans l'épreuve *vs* compétences complexes, comme la production écrite ou la compréhension, qui sont particulièrement difficiles à évaluer), la façon de constituer le barème et les seuils de réussite dans la plupart des disciplines. La manière de déterminer les seuils de réussite diffère également entre le primaire et le CO, ce qui a certainement des effets : dans un cas, il est déterminé a priori (en utilisant les prétests et une table de spécification), dans l'autre, il est défini sur la base des résultats des élèves (a posteriori) dans la plupart des disciplines (la physique constitue un exemple intéressant avec des seuils de réussite estimés au moment de la conception de l'épreuve en fonction des objectifs). Par ailleurs, l'ensemble des acteurs interrogés au CO ont relevé une difficulté majeure liée à la contrainte des mêmes objectifs pour tous. Comment concilier ce postulat avec l'existence des regroupements et des niveaux de compétences différents des élèves ? Faut-il avoir la même épreuve comme en français (branche sans niveau) avec des barèmes différenciés ou des épreuves un peu différentes comme en mathématiques ou en allemand (branches avec niveaux) ?. La plupart des personnes interrogées ont relevé ce paradoxe et les difficultés rencontrées de contenter tout le monde.

Utilité des épreuves

Les points forts

Pour la plupart des personnes interrogées dans les deux ordres d'enseignement, un des principaux points forts de ces évaluations est d'être commune à tous les élèves d'une cohorte. Cela permet d'avoir des exigences communes pour une discipline donnée. Elles permettent également de réguler l'enseignement et d'unifier les pratiques (voire parfois de les modifier par exemple en introduisant l'évaluation de la production orale en allemand au CO ou en introduisant des situations-problèmes en mathématiques). Elles peuvent aussi contribuer à une certaine crédibilité par rapport aux autres ordres d'enseignement (CO ou PO). Elles sont également importantes pour les élèves et les parents. Les enseignants du CO relèvent également que les épreuves permettent d'habituer les élèves à être évalués sur un champ plus large et de fonctionner de manière autonome.

De plus, les objectifs de ces épreuves définis institutionnellement sont confirmés par l'analyse de la qualité docimologique de quelques épreuves déjà administrées qui met en évidence une fiabilité satisfaisante des épreuves par discipline évaluée (français I, français II et mathématiques) pour deux utilisations : la certification (situer les élèves sur l'échelle du test ou par rapport à un seuil de réussite) et la régulation de l'enseignement.

---

[44] Précisons que la direction du CO cherche actuellement à généraliser les prétests.

[45] Le cadre de référence élaboré notamment pour le français et les mathématiques contribue à préciser le contenu de l'épreuve et serait une réponse à cette critique.

[46] Ce point faible est à nuancer puisque certaines personnes souhaitent s'engager sur le plus long terme. Par ailleurs, des formations ont été organisées pour les commissaires, notamment en langues.

Les points faibles

Si les deux fonctions, certification des élèves et régulation de l'enseignement qui figurent dans les textes officiels semblent confirmer dans les propos des acteurs interrogés et dans les analyses effectuées au niveau des épreuves, il ressort que ces épreuves sont utilisées pour répondre à trop de besoins : il y aurait notamment une confusion entre les objectifs visant à vérifier l'atteinte du programme par les élèves et le pilotage du système. Pour d'autres, il manque une évaluation système (notamment au primaire).

L'analyse et l'exploitation des résultats jugées insuffisantes sont soulignées par la majorité des acteurs.

Certains effets pervers des épreuves sont relevés notamment au CO : la tendance au bachotage et la réduction du programme et de l'enseignement au contenu des épreuves.

Certains enseignants du primaire déplorent la trop grande importance de ces épreuves aux yeux des élèves et de leurs parents, le cas particulier des épreuves de 2P trop lourdes et trop stressantes pour les élèves, le caractère normatif des épreuves cantonales en général. Le moment de l'année choisi est considéré comme étant trop tôt (le programme n'est pas terminé) pour les enseignants des deux ordres d'enseignement. Enfin, au CO, dans certaines disciplines, on estime que le type d'évaluation proposé est trop différent de ce qui se fait habituellement en classe (notamment pour le regroupement B).

Efficacité, efficience

Les points forts

Globalement l'analyse des parcours des élèves montre que les épreuves externes ont une bonne prédictivité même si ce n'est pas leur objectif prioritaire. En effet, une part très importante des élèves qui en 6P réussissent les épreuves cantonales se retrouvent encore en 9^{e} dans le regroupement A. L'analyse docimologique met également en évidence une bonne validité prédictive des épreuves de 6P sur celles de 7^{e}.

Au CO, un point particulièrement positif est l'existence d'un outil informatique très précieux, EVACOM qui permet la saisie des résultats aux épreuves cantonales ou communes (informations récoltées pouvant être les plus complètes possibles selon les disciplines) et leur consultation.

Les points faibles

La place des épreuves cantonales ou communes dans les carnets et leur prise en compte de manière plus globale est considérée comme un problème par l'ensemble des acteurs.

Au CO, certains enseignants relèvent la surcharge de travail due à la correction et à la saisie des résultats.

Propositions d'amélioration

En lien avec les points faibles relevés, les acteurs interrogés lors de la première phase de l'étude (autorités, directions de l'enseignement, responsables des épreuves et concepteurs) ont proposé un certain nombre d'améliorations qui se rapportent aux dimensions suivantes :

- une certaine standardisation des épreuves du point de vue de leur contenu, de leur organisation et la difficulté des items pour éviter les biais et permettre une comparabilité d'une année à l'autre ;
- une amélioration au niveau des conditions de passation (pour le primaire) et/ou de correction en croisant les classes notamment ;

- un développement de l'analyse des résultats et des prétests ainsi que des moyens ou ressources mis en œuvre pour la réaliser. L'outil informatique EVACOM, comme nous l'avons relevé à plusieurs reprises, est un outil précieux. Toutefois, il dépend des informations qui sont saisies. Si l'on veut en savoir plus sur les acquis des élèves et dans quels domaines se situent leurs lacunes, il est indispensable d'avoir des résultats par question comme c'est le cas par exemple pour les épreuves communes de mathématiques. Ce type d'analyse pourrait également faire l'objet d'un feed-back aux concepteurs de programmes.

Certaines questions restent en suspens :

- La question de la pluralité des finalités. On a pu constater qu'elles poursuivaient deux objectifs (évaluation des acquis des élèves et régulation de l'enseignement). A cela, s'ajoute parfois l'évaluation du système. D'autres auteurs ont montré (notamment, Davaud, Hexel, Bain, 1983 ; Monseur, Demeuse, 2005) que souvent les épreuves communes ou les évaluations externes poursuivent plusieurs buts ou donnent lieu à des utilisations variées. Doit-on faire jouer à la même épreuve tous ces rôles ? Ou devrait-on plutôt garder la fonction certificative et utiliser d'autres outils pour le pilotage (p. ex. futurs tests HarmoS et enquêtes internationales comme PISA). On peut supposer qu'avec l'introduction des directeurs d'écoles à l'école primaire et l'existence des directeurs d'établissement au CO, il est nécessaire d'avoir des outils pouvant donner des informations liées au fonctionnement des établissements.
- La question de la discrimination des items. On a pu constater dans l'analyse de la fiabilité que les items pourraient être mieux adaptés au niveau des élèves (ni trop faciles - cas de l'épreuve de 2P en compréhension de l'oral administrée en 2006, ni trop difficiles) pour davantage différencier les connaissances/compétences des élèves. Toutefois, on peut se demander ce que cela signifie au niveau des objectifs de l'épreuve. Au primaire, on cherche davantage à vérifier si les élèves ont atteint les objectifs fixés et non différencier les élèves. Au CO, la logique est un peu différente, on souhaite être informé sur la proportion d'élèves qui les auraient atteints et surtout à quel niveau se situent les élèves. Même si l'évaluation cantonale ou commune dans la scolarité obligatoire doit être plus cohérente d'un niveau d'enseignement à l'autre, on peut supposer que les objectifs du début du primaire par exemple et ceux de 9^{e} ne sont pas les mêmes. L'orientation prend une place certaine au CO et pourrait également en avoir une à la fin du primaire.
- La question des épreuves communes au CO. Cette question est très complexe. On peut décider de la remettre en discussion une fois la nouvelle organisation du CO votée. Désire-t-on des épreuves communes à tous les élèves dans toutes les disciplines, y compris celles à niveaux ou seulement dans celles sans niveau (se pose alors le problème de l'anglais) ? Une autre possibilité est celle du tronc commun : une partie commune qui représente par exemple les 2/3 de l'épreuve et une autre partie plus difficile pour les élèves de A et ceux de H avec un niveau comparable.
- Une autre question est le type d'évaluation et les différences entre disciplines. Plusieurs interlocuteurs ont évoqué le fait que les disciplines de type scientifiques avaient des objectifs clairs et semblaient plus faciles à évaluer même si tout ne peut pas être évalué dans une épreuve papier-crayon. Dans le domaine des langues, il semblerait qu'il soit plus difficile d'identifier des objectifs aussi spécifiques, comprendre ou écrire un texte fait davantage appel à des compétences globales ou complexes, plus compliquées à découper en micro-compétences (d'où la réticence des commissaires de français par rapport à la table de spécification).

Réflexions en vue d'une nouvelle organisation de l'évaluation cantonale et commune

Des différents éléments analysés ci-dessus, il ressort un tableau complexe d'où émergent peu de lignes directrices claires permettant de définir une organisation transversale de l'évaluation externe dans la scolarité obligatoire genevoise. Dans cette partie nous tenterons de faire ressortir un certain nombre de points pouvant servir de repères en vue d'une telle organisation future de l'évaluation externe à Genève. Ces points pourraient s'organiser autour des cinq pôles suivants qui sont bien sûr dépendants les uns des autres :

- objectifs et attentes des évaluations externes,
- contenu des épreuves et manière d'évaluer,
- procédures pour atteindre ces objectifs,
- administration, utilisation et exploitation de l'évaluation externe,
- compétences et ressources nécessaires.

Objectifs et attentes

Vérifier l'atteinte des objectifs et participer à la certification

Les informations recueillies montrent qu'il existe un consensus, en tout cas formel, sur les objectifs des épreuves externes genevoises aussi bien pour l'enseignement primaire que pour le CO. Elles visent avant tout la vérification de l'atteinte des objectifs du programme et elles entrent en ligne de compte pour la certification des élèves. Cependant dans la pratique, les épreuves sont parfois utilisées à d'autres fins (pilotage du système, p. ex.) pour lesquels elles n'ont pas été conçues. De plus, derrière cette cohérence formelle, il faut être attentif aux spécificités liées aux ordres d'enseignement, aux parcours scolaires et à l'âge des élèves : par exemple, utiliser une épreuve commune de type papier-crayon qui entre dans la certification pour des élèves de 2P. A cet âge, on peut se demander si l'aspect diagnostique ne devrait pas être privilégié. Pour les épreuves de 6P, si elles visent officiellement la vérification du programme, elles sont également utilisées pour l'orientation des élèves au CO. Pour le moment comme près de 75% des élèves sont orientés vers le regroupement A, le système fonctionne assez bien car la majorité des élèves réussissent les épreuves de 6P. On peut se demander ce qu'il en sera lorsque nous aurons un système qui regroupera les élèves en trois catégories. Il sera alors plus difficile d'utiliser les épreuves comme aide à l'orientation des élèves.

Épreuves communes au CO : un rôle d'orientation

Pour le CO, la question est encore plus complexe, car les épreuves mesurent l'atteinte des objectifs pour un même plan d'études avec des élèves qui sont placés dans des regroupements différents en fonction de leurs performances scolaires. Dès lors, des arrangements sont nécessaires pour réaliser des épreuves « communes » en termes de barème ou de contenu. Par ailleurs, du fait de l'existence de ces regroupements, les épreuves communes ont de fait également une fonction d'orientation et de sélection soulignée par les enseignants interrogés étant donné qu'elles y participent d'une certaine manière en donnant des indications sur le niveau des élèves.

Objectifs et finalités : la place des épreuves cantonales ou communes en perspective avec les autres évaluations externes

Dans la conjoncture actuelle, ces objectifs et attentes doivent être envisagées également dans une perspective régionale, en lien avec le projet de développement d'épreuves de références romandes, et même nationale, mise en place de standards de performance de base nationaux. Toutefois, pour l'instant, il est difficile de prendre en compte concrètement ces deux projets qui ne sont pas suffisamment avancés. Cependant, on peut esquisser quelques lignes possibles en fonction des éléments actuellement en notre possession. Le dispositif HarmoS se situe à un niveau monitorage du système. Les instruments développés dans ce contexte s'adresseront à des échantillons d'élèves pour vérifier si les différents systèmes présents en Suisse atteignent les standards nationaux de performance. Les épreuves de référence régionales qui, dans le projet de la CIIP, sont appelées « épreuves romandes communes » pourraient être des épreuves de même type que les épreuves cantonales actuelles car elles seront destinées à l'ensemble des élèves et mesureront l'atteinte des objectifs du programme d'études romand (PER). Comme ce programme devrait être en vigueur dans l'ensemble des cantons romands, logiquement les épreuves cantonales dans les domaines testés par les épreuves romandes devraient à terme disparaître si la finalité des épreuves de référence romandes (actuellement en discussion) est la certification. Dans le cas contraire, elles viendraient s'ajouter aux épreuves cantonales ou communes existantes.

Par ailleurs, la réflexion sur l'évaluation externe devrait également tenir compte de l'importance donnée à cette évaluation par rapport aux évaluations des élèves réalisées par les enseignants. Plus globalement, la place de l'évaluation par rapport au temps consacré à l'enseignement et l'apprentissage doit également faire partie de la réflexion plus générale sur les objectifs de l'école.

Contenu de l'épreuve et manière d'évaluer

Le format des questions et le choix des objectifs

D'autres éléments peuvent également faire l'objet d'un examen : le format des questions qui devrait être le plus varié possible comme c'est déjà le cas dans les épreuves existantes : QCM, questions fermées à réponse courte mais également questions ouvertes. Le format des questions est comme le souligne Pini et al. (2006) intimement lié aux types d'objectifs que l'on veut mesurer. En effet si l'on veut mesurer les compétences des élèves, il est important que l'élève crée sa réponse et ne se contente pas de la choisir parmi plusieurs propositions.

Ainsi, s'il est vrai que certaines questions à choix multiples ou fermées permettent une correction relativement facile et peu sujette à des biais de correction, on peut s'interroger sur une évaluation qui ne porterait que sur ce type de questionnement et éviterait aux élèves de construire eux-mêmes leurs réponses par souci de standardisation. La commission de mathématiques a mentionné cette question estimant que la suppression des problèmes ouverts pourrait être préjudiciable à la qualité de l'épreuve, appauvrissant l'évaluation des compétences des élèves. On prendra pour exemple l'enquête PISA ou les épreuves HarmoS qui ont permis de définir les standards de formation où l'on a cherché à multiplier les formes de questions (QCM, questions fermées à réponse courte mais également questions ouvertes qui supposent l'élaboration d'une réponse par les élèves et une codification soignée de la part des correcteurs). Monseur et Demeuse (2005) mentionnent également de nouvelles formes d'évaluation, *authentic assessment*[47] *alternative assessment*[48] permettant d'évaluer des processus cognitifs.

Par ailleurs, il serait aussi utile de déterminer le niveau de difficulté des items (p. ex. questions d'application *vs* situations-problèmes ou items de deux niveaux comme en mathématiques ou en physique).

La représentativité des objectifs du plan d'études

Un autre point crucial est le choix et la détermination de ces objectifs pour vérifier qu'ils soient bien représentatifs du plan d'études. A cela s'ajoute une autre contrainte : leur pondération dans l'épreuve. Les enseignants ont souvent relevé que les épreuves couvraient bien le programme prévu mais ont parfois émis des réserves quant à la représentation des objectifs dans l'épreuve.

La stabilité des épreuves

La stabilité des composantes prises en compte est également importante pour pouvoir comparer les épreuves d'une année à l'autre et pour garantir une certaine couverture du plan d'études (p. ex. composantes du français identiques d'une année à l'autre).

Mesurer des compétences complexes

Par ailleurs, la mesure des compétences complexes en langues est une question difficile à résoudre dans le contexte d'une évaluation papier-crayon en temps limité. Par exemple, on note une insatisfaction par rapport à l'évaluation de la production écrite. Au primaire, cette partie est bien réussie mais les critères utilisés ne vont pas suffisamment en profondeur pour analyser de près cette

[47] Il s'agit d'une évaluation visant à placer les élèves dans des situations ou des situations réelles.

[48] Contrairement à certaines évaluations traditionnelles principalement sous formes de QCM, elles cherchent plutôt à mettre l'élève dans une situation de création ou d'élaboration de réponse.

production. Au CO, la tâche de production est très limitée et on ne dispose pas d'une situation suffisamment riche pour évaluer les différentes facettes de la production écrite.

Procédures pour atteindre ces objectifs

Généralisation des prétests

Malgré la qualité technique globale des épreuves mentionnée ci-dessus, il apparaît indispensable pour améliorer la qualité des instruments que les prétests soient systématiquement mis en place auprès d'un nombre suffisant d'élèves (le nombre minimal d'élèves ou de classes devra être déterminé sur la base d'analyses). En effet, les prétests permettent de vérifier si le degré de difficulté des items est adéquat à la population visée, si les consignes sont claires, si les modalités de correction sont applicables et permettent d'évaluer les biais de correction. Ils sont également nécessaires pour fixer les seuils de réussite ou les confronter à ceux fixés a priori par les experts.

La réalisation de tels prétests implique une planification sur plus d'une année et suggère une stabilité des personnes engagées dans la confection des épreuves. L'analyse de ces prétests devrait être menée aussi bien de façon qualitative par des experts des domaines concernés que du point de vue statistique pour vérifier la qualité technique de l'épreuve (fiabilité p. ex.).

La prise en compte des différents niveaux de compétences des élèves (regroupements différenciés

Il reste cependant des spécificités selon les ordres d'enseignement dont il faudra tenir compte pour la réalisation des épreuves. Par exemple au CO, il s'agira de prendre en compte les différences entre le regroupement A et le regroupement B. Plusieurs approches peuvent être envisagées : des barèmes différents, c'est ce qui est actuellement appliqué de façon systématique. D'autres pistes pourraient être explorées. Tenir compte de la difficulté des items (p. ex. de la physique) pour estimer l'atteinte des objectifs du programme. Faire une partie commune pour les deux regroupements, c'est ce qui est fait pour les épreuves de mathématiques et de français de 9^{e} afin de les utiliser comme moyen de certification pour l'entrée en apprentissage, et des parties spécifiques pour chaque regroupement.

Administration, utilisation et exploitation

Conditions de passation et de correction

Une fois l'instrument créé, il s'agit de l'administrer, d'exploiter et d'utiliser les résultats. La question de l'administration et de la correction de l'épreuve est un des points pour lesquels nous avons eu le plus de commentaires lors du recueil des données. Il est vrai que l'administration et la correction des épreuves externes par les enseignants des élèves peuvent poser problème du point de vue de l'équité face à l'épreuve. Il est bien évidemment irréaliste de penser que ce soit possible d'organiser une administration et une correction externe de l'épreuve. Cependant, un certain nombre de mesures peuvent être envisagées. Certaines de ces mesures ont déjà été appliquées parfois de façon partielle et locale. Il s'agit par exemple d'échanger les copies entre enseignants de façon à ne pas corriger les copies de ses propres élèves. On peut envisager également une correction par établissement afin de diminuer les différences de correction entre enseignants.

Analyse des résultats

Les dispositifs mis en place pour saisir et fournir les informations nécessaires aux personnes qui doivent les recevoir, surtout enseignants et direction d'écoles, notamment afin de les intégrer aux résultats des élèves sont bien rôdées dans les deux ordres d'enseignement bien que fonctionnant sur des logiques différentes. Au primaire, on réalise une saisie centralisée sur la base d'une correction manuelle des enseignants (cette procédure est actuellement en train d'évoluer) alors qu'au CO, les données saisies directement par les enseignants dans EVACOM. On peut se demander si une harmonisation des dispositifs ne serait pas souhaitable. Par ailleurs, au CO, EVACOM est utilisé de façon assez différente selon les disciplines. Par exemple en mathématiques les données sont saisies au niveau de la question alors qu'en français on se contente de la somme des points par type de

compétences par partie de l'épreuve. Une réflexion devrait être menée sur le niveau d'information à saisir. Ce niveau dépendra du type d'information que l'on souhaite obtenir. Par exemple au niveau d'un prétest, il est indispensable d'avoir le détail des réponses au niveau de l'item afin de pouvoir estimer la qualité de ceux-ci. Si l'information est un nombre de points à convertir en note, à la limite le total des points obtenus à l'épreuve suffit.

On peut regretter actuellement que les données disponibles soient peu utilisées pour améliorer la qualité des épreuves notamment parce que le plus souvent on ne dispose pas d'information des niveaux des items, même pour un échantillon réduit d'élèves. Les quelques analyses docimologiques que nous avons menées montrent que la comparaison de classes ou d'établissements à l'aide des épreuves est complexe et devrait être menée avec prudence.

Compétences et ressources

Les compétences nécessaires : des équipes pluridisciplinaires

L'ensemble des éléments décrits dans les points précédents ne peut pas être réalisé sans un certain nombre de compétences et de ressources. Actuellement les compétences mises en œuvre relèvent essentiellement de deux domaines. Nous avons d'un côté des personnes qui connaissent bien les programmes et les questions d'enseignement. Ce sont ces personnes qui sont concrètement impliquées dans la réalisation des épreuves (p. ex. commissaires aux épreuves du CO et formateurs du primaire[49]). De l'autre côté, nous avons les responsables de l'évaluation commune et les directions des ordres d'enseignement qui ont une fonction de contrôle et de validation et qui en général devraient disposer des ressources méthodologiques et logistiques.

Toutefois, il faut être attentif au fait de ne pas multiplier les groupes et les structures mais plutôt développer des lieux où les personnes de compétences différentes ont l'occasion de travailler concrètement à une tâche commune : participer au processus de réalisation de l'évaluation externe. A nos yeux un développement des compétences méthodologiques en évaluation nous semblent un défi important pour tout système scolaire au moment où l'évaluation externe au niveau cantonal prend une importance croissante. De plus la maîtrise de ces compétences est encore plus importante au moment où se développent des évaluations de ce type au niveau régional et national. Cela permettra à Genève de participer activement aux synergies à créer entre les différents niveaux.

Des compétences multiples requises

Les procédures à mettre en place pour réaliser les épreuves externes qui tiennent compte des quelques éléments évoqués ci-dessus nécessitent une mise en synergie de ressources existantes, le développement de compétences nouvelles ceci dans une organisation cohérente à laquelle les principaux acteurs (autorités, directions de l'enseignement, concepteurs, enseignants) sont partie prenante. Les équipes des concepteurs d'épreuves devraient réunir plusieurs types de compétences : connaissance des programmes et de l'enseignement, de l'évaluation et de ses méthodes, du traitement des données et de leur exploitation. C'est ce type d'organisation qui par exemple est utilisé en France pour l'élaboration des épreuves nationales. Par ailleurs des groupes réunissant les principaux acteurs (autorités scolaires, concepteurs, commissions d'enseignants) devraient assurer la validation des épreuves un peu sur le modèle de ce qui existait dans l'enseignement primaire il y a quelques années.

Une organisation transversale ?

La question d'une structure transversale est à envisager avec prudence : si les compétences docimologiques peuvent être transversales aux deux niveaux d'enseignement, il nous semble que les concepteurs devraient se situer davantage au niveau de leur ordre d'enseignement étant donné les différences de plan d'études, ce qui ne signifie pas une absence de regard sur les autres degrés concernés. Une condition devrait être retenue : leur stabilité et un souci de formation en évaluation.

[49] Au départ, à l'école primaire, les enseignants (commissions) étaient également inclus dans le processus de consultation une fois les épreuves élaborées.

Les ressources nécessaires

Dans le cadre de ce travail, il avait été demandé aux deux ordres d'enseignement de fournir une estimation du coût des épreuves. Malheureusement, les informations obtenues sont de nature trop différente pour permettre une comparaison entre les deux ordres[50]. L'analyse du coût des épreuves n'a qu'un intérêt limité car les épreuves n'existent pas pour elles-mêmes mais elles sont un élément du dispositif d'évaluation des élèves qui lui-même n'est qu'un élément du temps de l'élève et de l'enseignant, l'objectif principal de l'école étant de faire acquérir aux élèves des connaissances et des compétences. Relevons que le coût déjà important et en augmentation au fil des années suite au nombre plus élevé d'épreuves risque de s'accroitre encore si l'on veut améliorer le dispositif : généralisation de prétests, analyse et exploitation des résultats, etc. Toutefois, de telles améliorations seraient bénéfiques et permettraient d'avoir des épreuves encore plus fiables, valides et prédictives. Leur prédictivité devrait par ailleurs être mise en perspective avec celle des évaluations réalisées par les enseignants et pour la fin de la 6P avec la préorientation.

L'ensemble de ces réflexions peut contribuer à cerner les principaux éléments à prendre en compte lors de la définition d'une nouvelle organisation et de sa mise en œuvre. Ces propositions devraient faire l'objet de discussions et déboucher sur un consensus entre les différentes parties prenantes.

Voici en guise de conclusion quelques éléments qu'il faudra également prendre en considération. Même si les évaluations cantonales ou communes sont davantage des épreuves mixtes puisqu'administrées et corrigées par les enseignants, on peut certainement trouver des pistes intéressantes dans les neuf caractéristiques proposées par Bishop (1995 et citées par Monseur et Demeuse, 2005) que devraient présenter une évaluation externe pour améliorer le rendement des élèves : 1) que l'épreuve ait des enjeux importants pour l'élève *(high stakes)*, c'est-à-dire que les résultats de l'élève à l'épreuve aient des conséquences sur sa certification ; 2) qu'elle soit critériée et pas normative ; 3) qu'elle se rapporte à une discipline spécifique ; 4) que les résultats transmis aux élèves comprennent plusieurs niveaux (ceci pourrait résoudre les problèmes des regroupements au CO) ; 5) qu'elle évalue un large éventail de savoirs enseignés à l'élève ; 6) qu'elle soit perçue comme équitable (étant donné ses conséquences) ; 7) qu'elle réponde aux exigences psychométriques pour ne pas être remise en question (à ce niveau, nos épreuves y répondent déjà mais certains paramètres pourraient être améliorés) ; 8) l'épreuve devrait évaluer la matière enseignée aux élèves (on n'est pas dans une évaluation de compétences de type PISA éloignée des contenus scolaires) ; 9) que l'épreuve s'adresse à une majorité des élèves.

Comme on peut l'observer, certaines de ces caractéristiques sont déjà remplies ou partiellement remplies : seules la première et la quatrième ne nous semblent pas ou partiellement appliquées. Pour la première, il est important de garder à l'esprit que l'on ne se trouve pas dans une situation d'examen. Ici, les évaluations concernées sont censées compléter celles des enseignants. Elles rentrent dans la certification mais étant leur aspect ponctuel, elles ne peuvent à elles seules remplir cette fonction. Il serait d'ailleurs intéressant de se donner les moyens de vérifier quel est le lien entre les résultats de l'année provenant des évaluations internes des enseignants et ces évaluations cantonales et communes.

Enfin, comme le soulignent Monseur et Demeuse, il est nécessaire de garder à l'esprit que les évaluations externes qui répondent à ces critères présentent des points positifs tels que l'égalisation des chances d'accès aux études, des « effets de reflux positif » (adaptation du curriculum aux finalités notamment de réussir les évaluations) et la comparabilité. Mais elles peuvent également avoir pour effet de réduire le curriculum *(teaching to test)* et d'inciter les enseignants au bachotage avec leurs élèves pour augmenter leurs performances.

[50] Étant donné les délais impartis pour réaliser cette étude, seule une estimation d'une partie des coûts, essentiellement le travail des concepteurs et des frais d'impression du matériel de test, a pu être réalisée. Une analyse plus détaillée en collaboration avec les services financiers fera l'objet d'un mandat complémentaire qui précisera les éléments souhaités.

Perspectives d'avenir

Cette étude a permis de mettre en évidence un certain nombre d'éléments dont on peut tenir compte pour améliorer le système actuel lors de la réorganisation de l'enseignement obligatoire. Il sera nécessaire d'obtenir l'adhésion des différents partenaires sans laquelle le nouveau dispositif ne pourra pas fonctionner. L'adhésion devra porter non seulement sur le dispositif mais également sur les objectifs et les démarches confiées à celui-ci. Le département devra également prendre une décision quant aux finalités des épreuves cantonales ou communes ainsi que leur place par rapport aux évaluations réalisées par les enseignants dans leur pratique régulière. Cette décision devra reposer sur un large consensus des différents acteurs de l'éducation. De plus l'ensemble du dispositif devra s'articuler avec les autres évaluations externes qui n'existent pas encore et dont les objectifs ne sont pas encore déterminés notamment en ce qui concerne les épreuves communes de référence romandes. Par ailleurs un suivi et une évaluation régulière du dispositif sera nécessaire afin d'en assurer la pertinence et l'efficacité.

Références bibliographiques

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations : la généralisabilité : mode d'emploi*. Genève : Centre de recherches psychopédagogiques.

Bain, D., Weiss, L., Agudelo, W. (à paraître). Radiographie d'une épreuve commune de mathématiques au moyen de la généralisabilité. In Actes du 20e colloque de l'ADMÉÉ-Europe 'Entre la régulation des apprentissages et le pilotage des systèmes ; Évaluations en tension'. Genève, 9-11 janvier 08.

Bertrand, R. & Blais, J.-G. (2004). *Modèles de mesure ; L'apport de la théorie des réponses aux items*. Québec : Presses de l'Université du Québec.

Bishop, J. H. (1995). The impact of Curriculum-Based External Examinations on School Priorities and Student Learning. *International Journal of Educational Research, 23, 8,* 653-752.

Bressoux, P. (1995). Les effets du contexte scolaire sur les acquisitions des élèves : effet-école et effets-classes en lecture, In *Revue française de sociologie, 36, 2*, pp. 273-294.

Cardinet, J. & Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

Davaud, C., Hexel, D., Bain, D. (1983). *Enquête sur les fonctions des épreuves communes auprès des Directeurs de Collèges et des Présidents de Groupes*. Document de travail. Genève : CRPP.

Diederonck, Ch. (2008). *Comment les évaluations externes des acquis des élèves sont-elles perçues par les enseignants du primaire dans les cantons de Neuchâtel, Vaud et Fribourg ?* Neuchâtel : Institut de recherche et de documentation pédagogique.

Gullickson, A.R. (2002). *The Student Evaluation Standards. How to Improve Evaluations of Students*. Thousand Oaks, CA : Corwin Press Inc.

IRDP (2007). *Pratiques cantonales concernant l'organisation d'épreuves, d'examens, de tests ainsi que de l'obtention de certificats ou diplômes au cours de la scolarité obligatoire. Tableaux comparatifs. Année scolaire 2007-2088.* Neuchâtel : Institution de recherche et de documentation pédagogique.

Monseur, Ch., Demeuse, M. (2005). Les évaluations externes permettent-elles une régulation efficace ? In Demeuse, M., Baye, A., Straeten, M.-H., Nicaise, J. et Matoul, A. (Eds). *Vers une école juste et efficace.* Bruxelles : De Boeck.

Moody, I. (2001). A case-study of the predictive validity and reliability of Key Stage 2 test results, and teacher assessments, as baseline data for target-setting and value-added at Key Stage 3. *The Curriculum Journal, 12(1),* 81-101.

Pini, G., Reith, E., Weiss, L., Bugniet, F. (2006). *Guide méthodologique pour l'évaluation et la mesure en éducation.* Genève : DIPCP (Développement et innovation pédagogique au Cycle d'orientation).

Ntamakiliro, L, Tessaro, W. (2002). La perception de l'évaluation externe des élèves par les enseignants primaires genevois. In Symposium *Épreuves externes : évaluation de la qualité des apprentissages ou des curricula*. Actes du colloque de l'Admééé, Lausanne 2002.

Scriven, M. Évaluation perspectives and procedure. In W.J. Popham (Ed.) *Evaluation in education*. Berkeley : McCutchan, 1974.

Shepard, L.A. (1997). A checklist for evaluating large-scale assessment programs. Paper 9. Occasional Paper Series. The Evaluation Center. Western Michigan University.

Soussi, A., Ducrey, F., Ferrez, E., Nidegger, Ch., Viry, G. (2006). *Pratiques dévaluation : ce qu'en disent les enseignants (à l'école obligatoire et dans l'enseignement postobligatoire général.* Genève : Service de la recherche en éducation.

Soussi, A., Petrucci, F., Ducrey, F., Nidegger, Ch. (2008). *Pratiques déclarées d'enseignement de la lecture et performances des élèves dans le canton de Genève.* Genève : Service de la recherche en éducation.

Stake, R.E. (1969) Evaluation, design, instrumentation, date collection, and analysis of data. In J.L. David (Ed). *Educational evaluation*. Colombus, OH : State Superintendent of Public Instruction.

Stufflebeam, D.L. (1974) Metaevaluation. Occasional Paper Series, no 3. Kalamazoo, MI : The Evaluation Center, Western Michigan University.

Stuffelbeam, D.L. (2003). Professional Standards and Principles for Evaluations. In *International Handbook of Educational Evaluation,* 279-302.

Wirthner, M., Ntamakiliro, L. (2008). Des épreuves de référence au service de l'évaluation des enseignants. In *Quelle évaluation des enseignants au service de l'école ? Actes du séminaire 2007 de l'AIDEP, Leysin, 6 et 7 décembre / organisés par Michel Guyaz*. Neuchâtel : IRDP (08.1).

Weiss, J. (éd.). (2008). *Quelle évaluation des enseignants au service de l'école ? Actes du séminaire 2007 de l'AIDEP, Leysin, 6 et 7 décembre / organisés par Michel Guyaz*. Neuchâtel : IRDP (08.1).

Annexes

Annexe 1. Liste des personnes interrogées

I^re^ phase (juin à septembre 2008) :

- le secrétaire général, M. Frédéric Wittwer ;
- le secrétaire adjoint, M. Renato Bortolotti ;
- les deux directrices de l'enseignement à la direction de l'enseignement primaire et du cycle d'orientation, Mmes Thérèse Guerrier et Bernadette Badoud-Volta ;
- le coordinateur des épreuves cantonales, adjoint à la direction de l'enseignement primaire, M. Ladislas Ntamakiliro ;
- le responsable du secteur de l'évaluation commune au cycle d'orientation (à la retraite depuis juillet 2008), M. François Bugniet ;
- le responsable de la préorientation au cycle d'orientation, M. Emiel Reith ;
- les concepteurs des épreuves du CEFEP :
  - en français : Mmes Muriel Wacker, Françoise Vodoz et M. Denis Métroz ;
  - en allemand : Mme Lotti Kuster ;
  - en mathématiques : Mmes Blandine Choquet et Muriel Corthésy, et MM. Eric Burdet et Jean-Pierre Bugnion ;
  - pour le SEDEV, Mme Christiane Jeannet ;
- les commissaires des épreuves du CO :
  - en français : Mmes Marianne Moser (PG), Emmanuelle Tarazzi et Caroline Salvi, MM. Dominique Pellizari et Bernard Pinget ;
  - en allemand : Mmes Christiane Winter (PG), Linda Souchard, Carole de Jong, Heike Vuillemin Beucher, Svetlana Stevanovic-Zaric et Gabriele Zimmermann ;
  - en mathématiques : Mmes Rita Joye-Bortolotti et Josiane Bloechlinger, MM. Claude Lecoultre (PG), Philippe Dubath (PG), Sébastien Archimède, Pierrick Dudognon et Andrea Reuben ;
  - en physique : MM. Christian Colongo et Jacques Bochet (PG).

II^e^ phase (décembre 2008-janvier 2009) :

- les enseignants des six écoles primaires sélectionnées : Crêts-de-Champel, Grottes, Hugo-de-Senger, Onex-Tattes, Palettes, Pâquis (35 enseignants en tout) ;
- les enseignants des cinq établissements du cycle d'orientation (Aubépine, Budé, Gradelle, Marais, Montbrillant) : en tout 76 enseignants du CO (allemand = 12 ; anglais = 12 ; biologie = 5 ; français = 24 ; mathématiques = 19 ; physique = 5).

Annexe 2. Le point de vue des autorités, des directions d'enseignement et des concepteurs

Grille d'entretien - Identification des points forts et faibles

Nous allons détailler les différents critères et vous demander de dire dans quelle mesure les affirmations suivantes vous semblent vraies *(tout à fait vrai, partiellement vrai, peu vrai, pas du tout vrai).*

	Tout à fait vrai	*Partielle-ment vrai*	*Peu vrai*	*Pas du tout vrai*
I. Adéquation ou qualité technique				
a) *Validité interne* L'évaluation commune/cantonale répond de manière univoque à la question à laquelle elle était supposée répondre.	4	9	--	--
b) *Validité interne* L'évaluation commune/cantonale mesure ce que l'on voulait mesurer. (NR=1)	5	6	1	--
c) *Validité externe* Les résultats de l'évaluation commune/cantonale ont la généralisabilité désirée (on peut les généraliser à d'autres populations, d'autres conditions de programmes ou d'autres moments avec fiabilité). (NR=1)	4	5	3	--
d) *Fidélité* Les informations récoltées sont précises et cohérentes. (NR=2)	1	8	2	
e) *Objectivité* L'évaluation est objective : d'autres évaluateurs arriveraient aux mêmes conclusions.	6	6	1	--
f) *Équité* L'évaluation est favorable ou défavorable de la même manière à tous les élèves (garçons/filles, francophones/allophones, élèves de culture et de milieux socioéconomiques différents). (NR=1)	6	4	2	--
II. Utilité				
g) *Pertinence* Les résultats sont pertinents pour les destinataires de l'évaluation. Chaque évaluation mène aux bonnes décisions pour l'élève. (NR=1)	4	8	--	--
h) *Importance* L'évaluation a pris en compte ce qu'il y a d'important et de significatif au niveau du programme. (NR=1)	10	2	--	--
i) *Étendue* L'information récoltée est suffisamment complète pour une évaluation qui réponde aux besoins des élèves.	6	4	3	--
j) *Crédibilité* Les différents acteurs ou publics considèrent l'évaluation comme valide et non biaisée.	9	2	2	--
k) *Qualification des évaluateurs* Les évaluateurs (personnes qui élaborent les épreuves) ont les compétences et la formation nécessaires pour produire des évaluations de qualité (épreuves). (NR=1)	4	7	1	--
l) *Qualification des évaluateurs* Les personnes chargées de l'analyse des résultats ont les compétences nécessaires pour produire des résultats utiles. (NR=5)	4	2	2	--
m) *Timing* Les résultats sont fournis aux acteurs concernés quand ils en ont besoin.	9	2	1	1

n) *Diffusion* Les résultats sont diffusés à tous les acteurs concernés (enseignants, élèves, parents, directions, etc.).	8	3	1	1
o) *Suivi* Les résultats fournis sont faciles à comprendre et informent clairement les différents acteurs sur la manière d'y donner suite ou d'assurer un suivi. (NR=2)	2	2	7	--
3. Efficacité et efficience				
p) *Efficience* L'évaluation est rentable compte tenu du temps et des ressources mises à disposition. (NR=1)	5	3	4	--
q) *Efficience* Compte tenu du temps et des ressources mises à disposition, l'évaluation atteint les objectifs fixés initialement. (NR=1)	6	6	--	--
r) *Efficacité* L'organisation actuelle (dégrèvement des enseignants) répond aux objectifs attendus. (NR=1)	4	5	3	--
s) *Efficience* Compte tenu du temps et des ressources utilisés, la prise en compte des EC dans l'évaluation des élèves est adéquate.	6	6	1	--
t) *Prédictivité* Les épreuves communes/cantonales sont un bon prédicteur de la réussite des élèves pour une discipline donnée. (NR=2)	2	9	--	--
u) *Lien avec les autres évaluations* Les résultats des élèves aux épreuves communes/cantonales sont cohérentes avec ceux des autres évaluations de l'année effectuées par les enseignants. (NR=2)	2	8	1	--

Annexe 3. Exemple de table de spécification en français II (6P)

Table de spécification de l'épreuve français II

objectif-noyau	composante	critères d'évaluation (objectifs spécifiques)	items	points	seuil de réussite - total -
OBSERVER LE FONCTIONNEMENT DE LA LANGUE	grammaire	• connaitre les types et formes de phrases	1	4	
		• connaitre les structures grammaticales	2	2	
			3	2	
		• connaitre les fonctions grammaticales	4	4	
		• connaitre les catégories grammaticales	5	3	
					10 /15
	orthographe	• savoir écrire sans faute	6	6	
		• connaitre les accords dans le GN	7	(4)	
			8	3	
		• accorder le verbe avec son sujet	7	4	
		• copier sans faute	7	1	
					9/14
	conjugaison	• s'exprimer en utilisant le temps des verbes qui convient	9	5	
		• connaitre les verbes et les temps de la liste du plan d'études et savoir utiliser le Roller	10	3	
			11	5	
					8/13
	vocabulaire	• reconnaitre les différents sens d'un mot selon le contexte d'emprunt	12	4	
			13	2	
		• comprendre une expression en contexte	14	3	
		• reconnaitre les mots d'une même famille	15	2	
		• établir un champ lexical	16	2	
		• former des mots à l'aide de la dérivation	17	3	
					11/16
		Total général :			**38/58**

Annexe 4. Questionnaire aux enseignants de l'école primaire et du CO concernant l'évaluation externe

Informations générales : enseignants primaires=36 ; enseignants CO=76 (allemand=12 ; anglais=12 ; biologie=5 ; français=24 ; mathématiques=19 ; physique=5)

1. Voici quelques affirmations sur l'évaluation externe (épreuves cantonales ou communes) qu'en pensez-vous ?

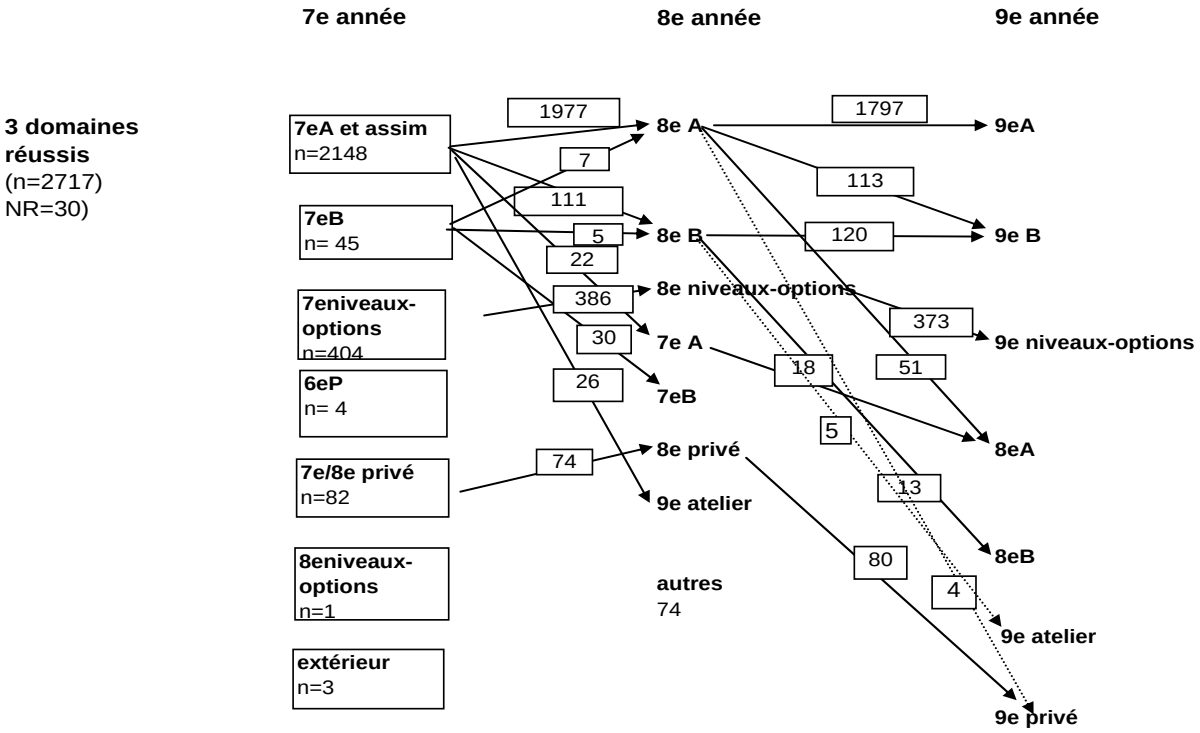
	tout à fait d'accord	*plutôt d'accord*	*plutôt pas d'accord*	*pas du tout d'accord*
a) L'évaluation externe peut être génératrice d'idées nouvelles (NR=2)	5.6 / 9.5	61.1 / 58.1	27.8 / 27.0	5.6 / 5.4
b) L'évaluation externe n'est pas pertinente car seul l'enseignant est à même d'évaluer les apprentissages de ses élèves parce qu'il les connaît bien	2.8 / 2.6	5.6 / 14.5	47.2 / 47.4	44.4 / 35.5
c) L'évaluation externe contribue à harmoniser les pratiques d'évaluation entre enseignants (NR=2)	2.8 / 29.7	77.8 / 41.9	13.9 / 21.6	5.6 / 6.8
d) L'évaluation externe permet une certaine objectivité parce qu'elle tient compte de ce qui est attendu pour un degré ou un cycle donné (NR=2)	22.2 / 20.3	61.1 / 48.6	11.1 / 23.0	5.6 / 8.1
e) L'évaluation externe permet d'évaluer et de réguler le système (NR=4)	5.7 / 16.4	60.0 / 53.4	22.9 / 23.3	11.4 / 6.8
f) Il est important d'avoir, à côté de l'évaluation régulière réalisée par l'enseignant, une évaluation commune à tous les élèves d'un degré ou pour un cycle car elle donne des repères précis et objectifs (NR=3)	31.4 / 33.8	51.4 / 45.9	11.4 / 16.2	5.7 / 4.1
g) L'évaluation externe sert à évaluer les enseignants (NR=3)	2.9 / 2.7	28.6 / 32.4	40.0 / 36.5	28.6 / 28.4
h) L'évaluation externe ne tient pas compte de ce qui est vraiment enseigné en classe	2.8 / 10.5	19.4 / 26.3	58.3 / 42.1	19.4 / 21.1
i) L'évaluation externe est utile pour s'assurer que les objectifs fondamentaux sont atteints (NR=1)	16.7 / 18.7	61.1 / 48.0	13.9 / 26.7	8.3 / 6.7
j) L'évaluation externe permet de savoir précisément ce qui est attendu (NR=7)	5.7 / 5.7	65.7 / 45.7	17.1 / 41.4	11.4 / 7.1
k) L'évaluation externe a pour effet de stresser les élèves (NR=1)	22.2 / 18.7	47.2 / 50.7	25.0 / 25.3	5.6 / 5.3
l) L'évaluation externe entraine un certain bachotage (NR=4)	8.8 / 27.0	47.1 / 58.1	32.4 / 13.5	11.8 / 1.4
m) L'évaluation externe oblige (ou a pour effet) de limiter le champ de l'enseignement au contenu des épreuves (ou de l'évaluation) (NR=2)	5.6 / 13.5	33.3 / 33.8	30.6 / 43.2	30.6 / 9.5

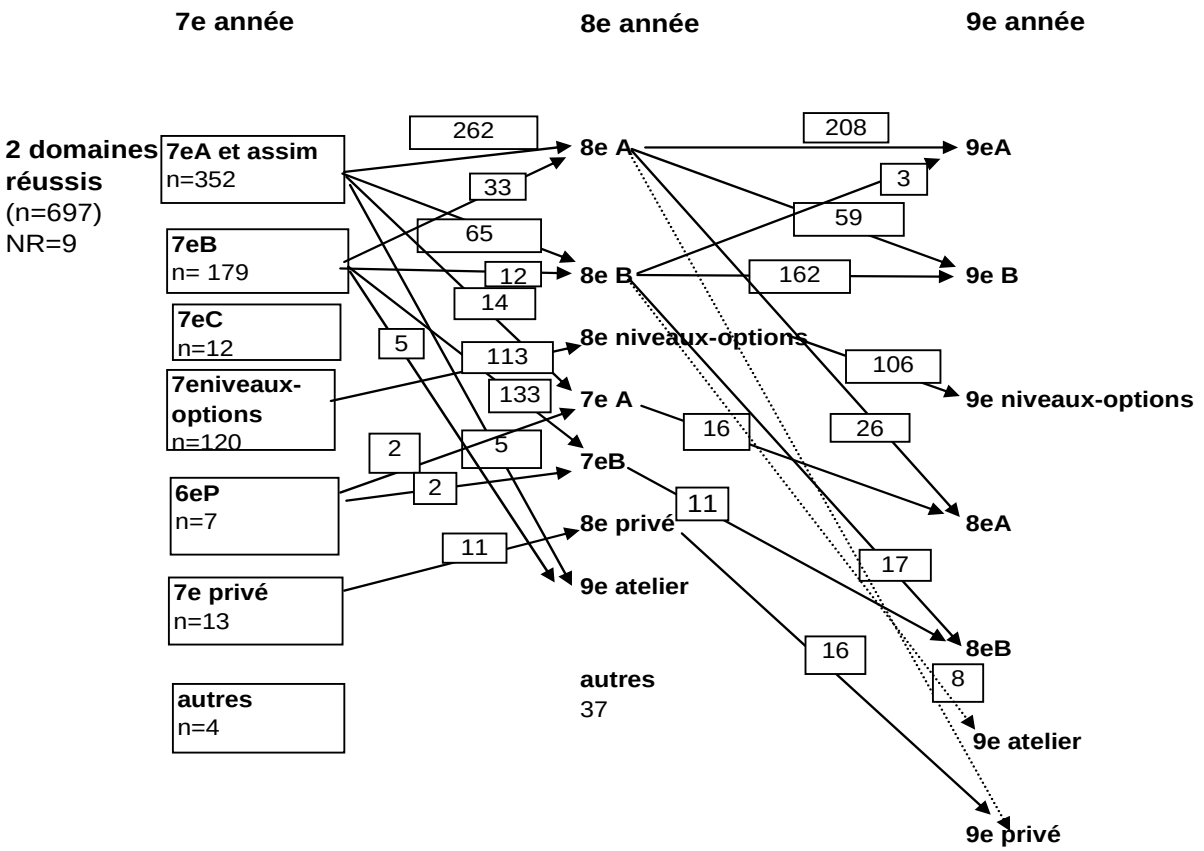
2. Identification des points forts et faibles

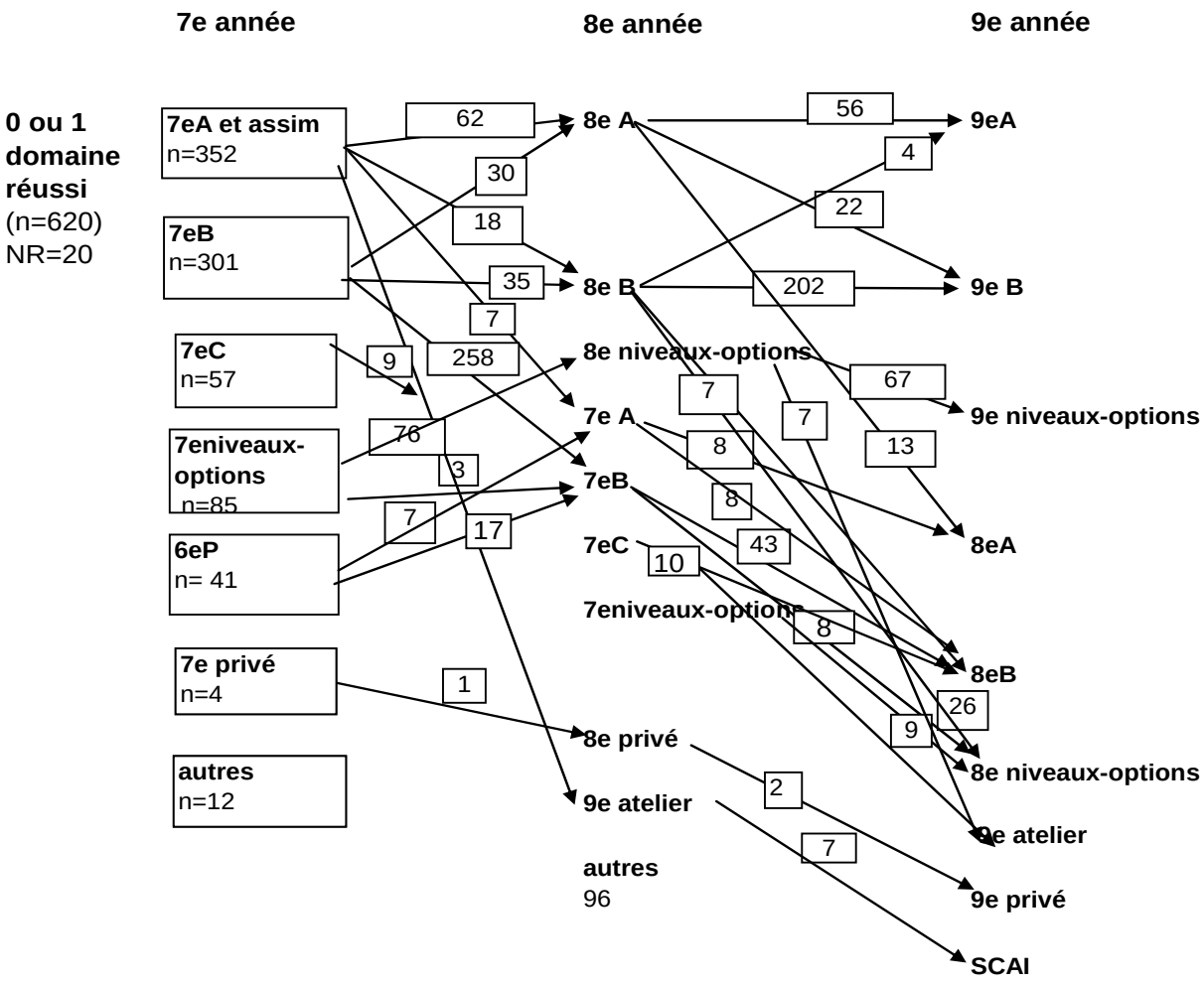
Veuillez indiquer dans quelle mesure les affirmations suivantes vous semblent vraies *(tout à fait vrai, partiellement vrai, peu vrai, pas du tout vrai).*

	Tout à fait vrai	*Partielle-ment vrai*	*Peu vrai*	*Pas du tout vrai*
I. Adéquation ou qualité technique (m = 1.64 / 1.54 / 1.57)				
a) L'évaluation cantonale ou commune constitue une mesure adéquate des acquis des élèves (NR=1)	2.8/8.0	80.6/60.0	11.1/26.7	5.6/5.3
b) Les informations récoltées sont précises et cohérentes (NR=3)	2.8/5.5	75.0/56.2	13.9/35.6	8.3/2.7
c) L'évaluation cantonale ou commune est objective : d'autres évaluateurs arriveraient aux mêmes conclusions (NR=4)	14.3/9.6	62.9/52.1	14.3/30.1	8.6/8.2
d) L'évaluation cantonale ou commune est favorable ou défavorable de la même manière à tous les élèves (garçons/filles, francophones/allophones, élèves de culture et de milieux socioéconomiques différents) (NR=5)	8.6/11.1	25.7/23.6	42.9/37.5	22.9/27.8
II. Utilité (m = 1.83/1.61/1.68)				
e) Les résultats sont pertinents pour les destinataires de l'évaluation (enseignants, parents, élèves, ….) (NR=1)	5.6/9.3	77.8/53.3	16.7/29.3	0.0/8.0
f) L'évaluation cantonale ou commune a pris en compte ce qu'il y a d'important et de significatif au niveau du programme	16.7/18.4	66.7/56.6	13.9/19.7	2.8/5.3
g) L'information récoltée est suffisamment complète pour une évaluation qui réponde aux besoins des élèves (NR=8)	2.9/11.6	60.0/42.0	37.1/39.1	0/7.2
h) L'évaluation cantonale ou commune est valide et non biaisée (NR=6)	16.7/14.3	58.3/44.3	16.7/32.9	8.3/8.6
i) Les résultats sont fournis aux acteurs concernés quand ils en ont besoin (NR=12)	37.5/19.1	43.8/38.2	15.6/23.5	3.1/19.1
j) Les résultats fournis sont faciles à comprendre et informent clairement les différents acteurs sur la manière d'y donner suite ou d'assurer un suivi (NR=7)	5.9/12.7	47.1/31.0	41.2/45.1	5.9/11.3
III. Efficacité et efficience (m = 1.79/1.62/1.67)				
k) Compte tenu du temps et des ressources utilisés, la prise en compte des EC dans l'évaluation des élèves est adéquate (NR=3)	5.7/20.3	65.7/45.9	22.9/28.4	5.7/5.4
l) Les épreuves communes/cantonales sont un bon prédicteur de la réussite des élèves pour une discipline donnée (NR=4)	5.7/6.8	60.0/42.5	31.4/39.7	2.9/11.0
m) Les résultats des élèves aux épreuves communes/cantonales sont cohérentes avec ceux des autres évaluations de l'année effectuées par les enseignants (NR=2)	20.0/8.0	62.9/52.0	14.3/33.3	2.9/6.7

Annexe 5. Parcours d'élèves de la 7e à la 9e en fonction de leurs résultats aux EC de 6P




7e année
8e année
9e année
2 domaines réussis
(n=697)
NR=9
7eA et assim
n=352
7eB
n= 179
7eC
n=12
7eniveaux-options
n=120
6eP
n=7
7e privé
n=13
autres
n=4
262
33
65
12
14
5
113
133
5
2
2
11
8e A
8e B
8e niveaux-options
7e A
7eB
8e privé
9e atelier
autres
37
208
3
59
162
106
16
26
11
17
16
8
9eA
9e B
9e niveaux-options
8eA
8eB
9e atelier
9e privé


7e année
8e année
9e année
0 ou 1 domaine réussi (n=620) NR=20
7eA et assim n=352
7eB n=301
7eC n=57
7eniveaux-options n=85
6eP n= 41
7e privé n=4
autres n=12
8e A
8e B
8e niveaux-options
7e A
7eB
7eC
7eniveaux-options
8e privé
9e atelier
autres 96
9eA
9e B
9e niveaux-options
8eA
8eB
9e niveaux-options
9e atelier
9e privé
SCAI
62
30
18
35
7
9
258
76
3
7
17
1
56
4
22
202
67
7
7
8
13
8
43
10
8
26
9
2
7

Annexe 6. Tableau récapitulatif des différents types d'épreuves (cantonales/ communes, romandes et nationales)

Épreuves communes ou cantonales

A l'école primaire

<table>
<tr><th rowspan="2">Degrés</th><th rowspan="2">Objectifs</th><th rowspan="2">Contenus</th><th colspan="3">Modalités</th><th rowspan="2">Utilisation des évaluations</th></tr>
<tr><th>Conception</th><th>Réalisation</th><th>Traitement et analyse</th></tr>
<tr><td>2P</td><td rowspan="3">Évaluation certificative en fin de cycle

Vérifier l'atteinte des objectifs</td><td>- français I (compréhension de l'écrit, production écrite et éventuellement compréhension de l'oral)
- français II (à titre informatif)
- maths
- écriture-graphisme</td><td rowspan="3">Formateurs des services de la DGEP (langues et mathématiques) sur la base du plan d'études (table de spécification des objectifs)</td><td rowspan="3">Essais dans les classes en début d'année dans le degré supérieur (3,5 ou 7)

Barème a priori à la suite des essais (seuil de réussite à environ 2/3 des pts)</td><td rowspan="3">- résultats pour l'ensemble des classes et des élèves
- résultats par classe
- résultats par école</td><td>- participe à la certification des élèves et au passage dans le degré ou le cycle suivant</td></tr>
<tr><td>4P</td><td>- français I compréhension de l'écrit, production écrite et éventuellement compréhension de l'oral)
- français II
- mathématiques
- allemand</td><td>- participe à la certification des élèves et au passage dans le degré ou le cycle suivant</td></tr>
<tr><td>6P</td><td>- français I compréhension de l'écrit, production écrite et éventuellement compréhension de l'oral)
- français II
- mathématiques
- allemand</td><td>- participe à la certification des élèves et au passage au CO
- français I et II, mathématiques : rôle d'orientation (avec les notes de l'année) par rapport aux regroupements A ou B (note 4 dans les 3 domaines pour A)</td></tr>
</table>

Au CO

Degrés	Objectifs	Contenus	Modalités			Utilisation des évaluations
			Conception	*Réalisation*	*Traitement et analyse*	
7e	Évaluation certificative (fin d'année) Vérifier l'atteinte des objectifs	- français - maths - allemand - latin	Commissions par discipline responsables des épreuves communes (enseignants sous la responsabilité d'un-e PG) sur la base du plan d'études de la discipline	*Essais dans les classes en début d'année dans le degré supérieur* Barème a posteriori variant selon les regroupements : A (2/3 environ des pts, H autour 60% et B 51%)	Enseignants entrent les résultats des élèves sur EVACOM *- résultats pour l'ensemble des classes et des élèves* *- résultats par classe* *- résultats par école*	- participe à la certification des élèves et au passage dans le degré ou le cycle suivant
8e		- français - maths - allemand - anglais - latin - physique - biologie				- participe à la certification des élèves et au passage dans le degré ou le cycle suivant
9e		- français - maths - allemand - anglais - latin - physique				- participe à la certification des élèves et au passage dans les différentes formations du PO

HarmoS - tests de référence (projet)

Degrés	Objectifs	Contenus	Modalités			Utilisation des évaluations
			Conception	*Réalisation*	*Traitement et analyse*	
2e (4e)	Évaluation et pilotage du système Vérifier l'atteinte des standards de performances	- Langue d'enseignement - Langues II et III - Mathématiques - Sciences	Experts nationaux ou institution mandatée ?	?	?	Vérifier l'atteinte des objectifs fixés par les standards minimaux définis au plan national notamment au niveau national, cantons et écoles
6e (8e)						
9e (11e)						

Convention scolaire romande - épreuves de référence (projet)

Degrés	Objectifs	Contenus	Modalités			Utilisation des évaluations
			Conception	*Réalisation*	*Traitement et analyse*	
2e ?	Évaluation certificative (fin d'année) ? Compléter l'approche HarmoS (spécificités romandes sur la base du Plan d'études romand) Vérifier l'atteinte des objectifs du PER	?	?	?	?	Vérifier l'atteinte des objectifs en fonction du PER Test de référence (HarmoS) pourrait servir d'épreuve commune romande
6e						
9e						

Enquête PISA

Degrés	Objectifs	Contenus	Modalités			Utilisation des évaluations
			Conception	*Réalisation*	*Traitement et analyse*	
9e	Évaluation des compétences de jeunes de 15 ans (+ 9e en Suisse)	- lecture (littératie) - mathématiques (culture mathématique) - sciences (culture scientifique	Consortium scientifique international chargé de l'enquête (également items proposés par les pays). Participation active des pays à toutes les phases du dévelop-pement des instruments à l'analyse des données	Prétest dans tous les pays participants une année avant le test principal Administration par des personnes externes à l'école Exigence sur un taux de réponse minimal des pays et des écoles participantes	- base de données internationale publique - résultats par pays et en fonction du contexte économique, culturel et scolaire des élèves - résultats sans identification d'élèves ou d'écoles - pas de résultat individuel possible	Évaluation de système permettant la comparaison des pays et des régions en fonction de différentes sous-populations

Département de l'instruction publique
Service de la recherche en éducation
12, Quai du Rhône ● 1205 Genève
Tél. +41/0 22 546 71 00 ● Fax +41/0 22 546 71 02 ● *prénom.nom*@etat.ge.ch ● www.ge.ch/sred

Papier recyclé