Note d'information du SRED
Numéro 54
Décembre 2012

Service de la recherche en éducation
Département de l'instruction publique, de la culture et du sport

Les évaluations externes dans quelques pays ou régions d'Europe : élaboration, analyses et diffusion

Anne Soussi et Christian Nidegger

Cette note complète le rapport *Évaluation des acquis à l'école obligatoire* et se centre sur les évaluations externes dans quelques pays ou régions d'Europe.

Par évaluation externe, nous entendons toute évaluation qui n'est pas réalisée par les enseignants de l'école et ce quelle que soit sa fonction (monitorage, bilan, évaluation certificative ou sommative, évaluation formative).

L'évaluation est un phénomène complexe de manière générale et dans le domaine de l'éducation en particulier. Elle suppose globalement un jugement orienté vers un but et qui va donner lieu à une prise de décision.

Les évaluations peuvent viser différentes fonctions et être internes comme externes. Elles peuvent également se situer à différents niveaux. Weiss (2002) distingue, quant à lui, 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 **(Tableau 1)**. 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).

Dans cette note, il sera question des évaluations externes, qu'elles suivent la logique de l'apprenant ou de l'organisation. Par contre, on n'abordera pas l'enquête internationale PISA étant donné que l'on s'intéresse ici aux dispositifs mis en place dans les différents pays (ou régions) sélectionnés. Depuis les années 1990, les évaluations externes se sont considérablement développées en Europe sur le modèle de l'Amérique du Nord. D'après Mons (2009), elles se sont transformées au fil des années: centrées au départ sur la mesure des apprentissages des élèves, elles ont des visées plus larges mêlant le pédagogique et le politique, servant d'outil de pilotage. Leurs effets peuvent être considérés par les différents acteurs comme positifs (harmonisation des pratiques pédagogiques et évaluatives, clarté des objectifs, etc.) mais également négatifs (tendance au *teaching to test*, stress des élèves et des enseignants, choix d'objectifs cognitifs au détriment de compétences sociales, types de questionnement et de compétences mesurées restreints, etc.).

Dans certaines analyses, à la fonction ou objectif s'ajoute le rôle, l'importance

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	Évaluation bilan des systèmes de formation (cantons, pays)
	Épreuves bilan pour la certification des apprentissages des élèves	Évaluation pour la certification des organisations apprenantes (école, établissement)
Évaluations internes	Évaluation formative des apprentissages des élèves	Évaluation régulatrice autogérée d'une organisation apprenante
	Évaluation bilan-sommative des apprentissages des élèves	

Nous tenons à remercier les représentants des pays concernés qui ont bien voulu répondre à ce sondage et sans qui cette étude n'aurait pas été possible.

ou l'impact qu'on attribue aux résultats des évaluations: enjeux élevés ou faibles (*high* ou *low stakes*). Ainsi, Ntamakiliro et Tessaro (2010) distinguent deux types d'évaluations externes, les standardisées et les non standardisées (notamment sur le plan des conditions de passation et de correction). Sur le modèle nord-américain, les épreuves externes peuvent encore se différencier en deux catégories: celles qui présentent pour les élèves des enjeux élevés et celles avec des enjeux faibles, selon "l'importance accordée aux résultats à ces épreuves dans les procédures de certification ou d'orientation scolaire" (p. 3). Mons distingue deux sortes de "rendre compte", l'*accountability dure*, modèle anglo-saxon, qui aurait pour conséquence des sanctions ou des récompenses à différents niveaux (écoles, enseignants ou élèves) et l'*accountability douce*, modèle européen, qui n'aurait pas de telles répercussions.

L'évaluation externe standardisée en Europe

Avant de nous centrer sur le sondage réalisé dans quelques pays ou régions d'Europe et au Canada, nous donnerons quelques éléments de la situation européenne concernant l'évaluation externe. Le rapport Eurydice dont émanent les informations qui suivent traite des évaluations standardisées ou tests nationaux. Il s'agit globalement des évaluations centralisées, nationales ou régionales. Dans ce rapport, trois groupes de tests nationaux sont distingués: ceux qui ont pour fonction le bilan des acquis des élèves (tests sommatifs), ceux qui servent à piloter et évaluer les établissements et/ou le système éducatif, et enfin ceux qui ont pour but de "contribuer au processus d'apprentissage des élèves à titre individuel en identifiant leurs besoins d'apprentissage spécifiques et en adaptant l'enseignement en conséquence" (2009, p. 8), c'est-à-dire une évaluation de type formative. Les auteurs précisent qu'il s'agit des principaux objectifs, certaines évaluations poursuivant plusieurs objectifs.

Comme on peut l'observer **(Tableau 2)**, on peut distinguer deux types de pays: ceux (les plus nombreux) qui possèdent, à côté de l'évaluation interne pratiquée par les enseignants, une évaluation externe sous différentes formes, et ceux qui n'en avaient pas (en tout cas au moment de l'enquête). Relevons que la plupart des pays possèdent une évaluation de type monitorage et bon nombre une évaluation de type sommative. L'évaluation de type formative ou diagnostique s'avère plus rare, sans doute parce que ce type d'évaluation est davantage le fait des enseignants eux-mêmes. On peut également souligner que plusieurs pays n'ont (ou n'avaient) aucune évaluation standardisée en 2008-2009: la Communauté germanophone de Belgique, la République tchèque, la Grèce, le Pays de Galles ou encore le Liechtenstein. Ceci n'est pas très étonnant pour de petites entités comme la Communauté germanophone de Belgique ou le Liechtenstein.

Les évaluations du premier groupe ont donc pour but de prendre des décisions concernant l'orientation ou le passage d'une classe à l'autre: 17 pays possèdent des évaluations de ce type. Dans la plupart des cas, ces tests sont obligatoires et sont organisés à la fin de l'enseignement secondaire I, c'est-à-dire le plus souvent à la fin de l'école obligatoire. Quelques pays ont des tests à enjeux importants à la fin du primaire comme la *Communauté française de Belgique*[1] ou la Pologne.

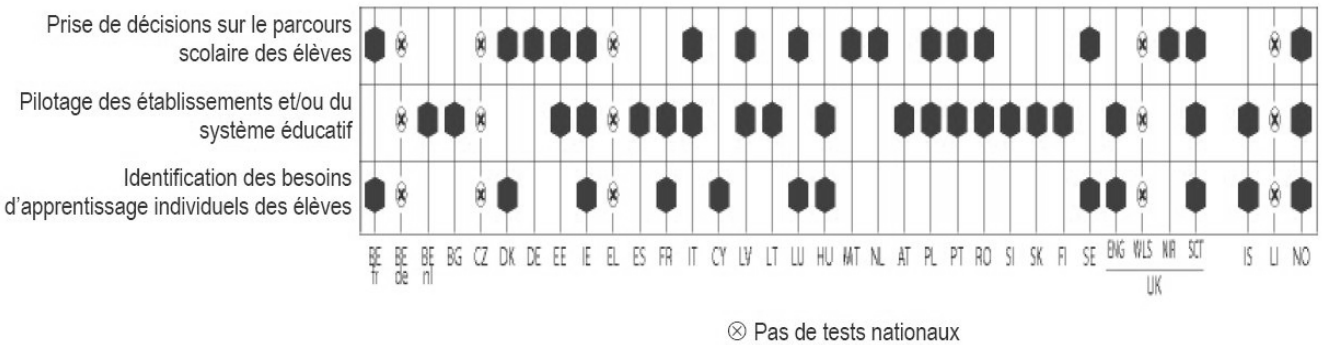
La seconde fonction, pilotage ou évaluation des établissements, concerne encore plus de pays (21). Il peut s'agir de comparaison entre établissements tout comme de l'évaluation de l'ensemble du système. Ces évaluations servent comme indicateurs de la qualité de l'enseignement ou de l'efficacité globale des politiques éducatives. Certains pays mettent l'accent sur l'évaluation au niveau des établissements (Lettonie, Hongrie, Autriche et Angleterre) tandis que d'autres le portent plutôt sur le niveau systémique (Belgique flamande, Estonie, *Irlande*, Espagne, *France,* Lituanie, Roumanie, Finlande et *Écosse*).

La troisième fonction, plus formative, s'observe dans seulement 12 pays ou régions. Enfin, trois pays possèdent des évaluations des trois types: *Irlande* (Eire), *Écosse* et Norvège.

Les moments auxquels ont lieu les évaluations varient beaucoup d'un pays à l'autre. La majorité des pays organisent les tests plutôt à la fin du primaire ou du secondaire I, mais on en trouve également qui le font plutôt au début la scolarité en 2e ou 3e année (Danemark, Estonie, *France*, *Italie*, Lettonie, *Luxembourg,* Suède, *Royaume-Uni* [Angleterre et Écosse] et Norvège).

Les matières concernées sont avant tout la lecture ou la production d'écrits et les mathématiques, surtout au primaire. Cet éventail est plus développé au secondaire, quand les évaluations sont certificatives et administrées à la fin du secondaire inférieur. Certains pays limitent leurs évaluations à deux ou trois matières, le plus souvent la langue d'enseignement et les mathématiques, voire une langue étrangère (*Italie*, Chypre, Portugal et Slovaquie; Allemagne, Autriche, Slovénie, Islande, Norvège et *Luxembourg*) tandis que d'autres les font porter sur plus de disciplines avec une rotation selon les années. Les sciences viennent également s'ajouter à ces disciplines. Notons que quelques pays évaluent aussi parfois les capacités transversales, comme en Écosse la résolution de problème, le travail en équipe et les TIC, ou dans d'autres cas des domaines interdisciplinaires comme en Communauté flamande de Belgique ou en Pologne.

Tableau 2. Principaux objectifs des tests nationaux standardisés - Niveaux CITE 1 et 2, 2008/2009



Source: Eurydice (2009, p. 25)

Tableau 3. Évaluations externes en Suisse romande

	Évaluations à visée sommative									Évaluations à visée diagnostique								
	3e	4e	5e	6e	7e	8e	9e	10e	11e	3e	4e	5e	6e	7e	8e	9e	10e	11e
BE															X			
FR						X			X		X		X					
GE		X		X		X	X	X	X									
JU						X							X				X	
NE						X		X		X	X	X	X	X			X*	
VS				X		X		X	X		X*	X*	X*	X*	X*			X*
VD		X		X		X			X**								X	
Total		2		3		6	1	3	4	1	3	2	4	2	2		3	1

Les cases grisées représentent les épreuves inscrites dans la loi scolaire du canton.

* Il s'agit d'épreuves à disposition des enseignants avec une passation libre. – ** Épreuve d'établissement utilisée dans le cadre de la certification de l'élève.

L'évaluation externe en Suisse romande

Au moment où le nouveau plan d'étude romand (PER) entre en vigueur et qu'une réflexion est menée autour d'une évaluation romande des acquis des élèves, il est intéressant de se pencher sur les évaluations cantonales existantes en Suisse romande. En 2009-2010, c'est-à-dire deux ans avant la mise en place du PER, il existait des situations très variées dans les cantons, certains préférant une évaluation diagnostique formative dans les petits degrés, d'autres ayant choisi une évaluation sommative. Le **Tableau 3** illustre cette grande diversité[2] qui rend la réflexion sur une évaluation romande complexe.

Relevons tout d'abord qu'à part Genève et dans une moindre mesure Vaud, la plupart des cantons possèdent les deux types d'évaluation externe (sommative et diagnostique/formative). Le canton de Berne se situe à part avec une seule évaluation en 8P (ex-6e). Comme dans l'étude européenne, les disciplines évaluées sont le plus souvent la langue d'enseignement et les mathématiques au primaire avec parfois une seconde langue dès la 6P (ex-4e). Au secondaire s'ajoutent souvent d'autres disciplines, notamment les sciences. Il ne semble pas y avoir, comme dans les pays ou régions européennes, d'évaluation des compétences transversales.

De manière globale, ce sont aux deux moments de transition (8P et 11e) que l'on trouve le plus grand nombre d'évaluation de type sommatif, ce qui n'a rien d'étonnant puisque l'on se trouve à la fin d'un cycle. En reprenant la distinction entre évaluations à enjeux faibles et élevés, Ntamakiliro et Tessaro soulignent la variation au niveau de ces enjeux selon les cantons. Ils seraient faibles à Fribourg et dans le canton de Vaud mais élevés au Jura, à Genève, à Neuchâtel et en Valais. Cette distinction ne porte que sur les épreuves standardisées, c'est-à-dire plutôt les évaluations sommatives. Précisons toutefois que la notion d'enjeux élevés doit être relativisée étant donné que les répercussions ne sont pas directes (les résultats des épreuves participent à la certification sans être déterminants) ni du même type qu'en Amérique du Nord.

Sondage auprès de quelques pays et régions caractéristiques

Afin d'en savoir plus sur la manière de procéder dans quelques pays européens et au Canada, nous avons réalisé un sondage succinct à propos de différentes questions techniques telles que l'élaboration des évaluations, leur contenu, les prétests et leurs analyses, la passation et la correction, l'analyse et la diffusion des résultats et la régulation des évaluations[3].

Le sondage ne se veut pas exhaustif même pour un pays donné. En effet, dans certains cas, il existe un organisme qui réalise toutes les évaluations externes tandis que dans d'autres, selon les fonctions (certificative, monitorage, etc.), elles sont réalisées par des organismes ou personnes différents, ce qui a rendu plus difficile la récolte d'informations.

Les pays interrogés ou régions sont au nombre de 9: la Communauté française de Belgique, la France, l'Écosse et le Royaume-Uni dans son ensemble, les Pays-Bas, l'Italie, le Luxembourg, la Suisse alémanique, et pour l'Amérique du Nord, l'Ontario. La plupart des pays ou régions sélectionnés l'ont été parce qu'ils possédaient une certaine tradition en matière d'évaluation externe. Avant de donner les résultats de cette enquête, nous allons décrire brièvement l'évaluation externe présente dans les pays ou régions concernés **(Tableau 4)**.

Objectifs et fonctions de l'évaluation

Nous nous centrons ici sur les évaluations externes, qu'elles se focalisent sur l'apprenant ou l'organisation. Nous avons retenu au départ quatre objectifs différents: 1) *évaluation de système* ou *monitorage*, 2) *bilan, 3) certification des élèves* ou encore 4) *diagnostique et/ou formative.*

Les fonctions les plus fréquemment attribuées aux évaluations externes sont la certification des élèves et l'évaluation de systèmes ou le monitorage. Les évaluations diagnostiques ou formatives sont (dans la plupart des cas, la France constituant une exception) des évaluations internes réalisées par les enseignants.

Les évaluations-bilans sont un peu plus rares et parfois difficiles à différencier de la certification des élèves ou du monitorage. On peut supposer que certaines évaluations poursuivent deux buts. C'est pourquoi nous parlerons surtout de trois fonctions: certificative, monitorage / pilotage du système, formative / diagnostique **(Tableau 5)**.

De manière globale, les contenus des évaluations sont le plus souvent la langue d'enseignement et les mathématiques, auxquels s'ajoutent, plutôt dans le secondaire, les langues 2, les sciences ainsi que plus rarement les sciences humaines (histoire et géographie, voire éducation citoyenne).

Si l'on essaie de dégager des tendances, il semblerait que l'évaluation certificative ait lieu par définition en fin de cycle (parfois en début) et porte sur les disciplines principales. Elles concernent tous les élèves d'un degré.

Les évaluations de type pilotage ou monitorage procèdent souvent par échantillon d'élèves ou d'établissements et ont lieu plus ou moins aux mêmes moments que les évaluations certificatives:

Tableau 4. Évaluations externes dans quelques pays ou régions caractéristiques

Pays ou régions	Objectifs des tests nationaux	Moments de la scolarité	Matières évaluées
Communauté française de Belgique	Évaluation certificative	6e année primaire et 1re année secondaire	Français, formation mathématique, éveil-initiation scientifique et éveil-formation historique et géographique
	Évaluation formative / diagnostique	- 2e (7a), 5e (10°) - 2e secondaire (13a)	Dépend des années (par cycle triennal): sciences, histoire et géographie; lecture, production écrite et langues 2 (6e); mathématiques et langues 2 (2e sec.)
France	Évaluation formative / diagnostique	3e année primaire (CE2) (facultatif); 1re année secondaire I (6e) (obligatoire)	Français et mathématiques
	Pilotage des établissements ou du système éducatif	- Échantillons - Fin de primaire (10-11a) - Fin de scolarité obligatoire (14-15a)	- Test national 1: Évaluations-bilan de toutes les matières enseignées sauf arts et sports par rotation (cycle de 5 a) - Test national 2: Évaluations-bilan des compétences de base en français et en mathématiques en fin d'école et de collège: français et mathématiques
Italie	Évaluation certificative	3e année secondaire I	Italien et mathématiques Dès 2010/2011: sciences et anglais en plus
	Pilotage des établissements ou du système éducatif	2e et 5e années primaire; 1re année secondaire I	
Luxembourg	Évaluation certificative	6e année de primaire (11a)	Allemand, français et mathématiques
	Évaluation formative / diagnostique	- 3e (9a) - 5e année sec. (15) en début d'année	- Allemand, mathématiques - Français pour les élèves de 5e secondaire
Angleterre	Évaluation formative / diagnostique	3e, 4e, 7e et 8e années facultatifs (8, 9, 10, 12, 13 ans) mais utilisés par la majorité des établissements	Anglais et mathématiques
	Pilotage des établissements ou du système éducatif	- 2e (7a, fin du Key Stage 1) - 6e (11a, fin du Key Stage 2)	Key Stage 1: anglais et mathématiques Key Stage 2: anglais, mathématiques et science
Écosse	Évaluation certificative	Facultatif mais presque tous les élèves des écoles publiques: 3e et 4e années du secondaire	National Qualifications (NQ); standard grade ou Intermediate Toutes les matières (7 ou 8) dont l'anglais et les mathématiques
	Évaluation formative / diagnostique	5 tests banque nationale (5-14 ans) facultatif	National 5-14 Assessment Bank : langue maternelle (anglais ou gaélique) et mathématiques
	Pilotage des établissements ou du système éducatif	3e, 5e, 7e années primaire et 2e année secondaire (8, 10, 12, 14 ans)	Scottish Survey of Achievement (SSA) obligatoire: langue maternelle (anglais ou gaélique), mathématiques, sciences, sciences sociales (centration sur un domaine chaque année)
Irlande du Nord	Évaluation certificative	Fin du Key Stage 3 facultatif	Anglais, mathématiques et science & technologie
Irlande	Évaluation certificative	Fin de 3e année post-primaire	Matières de base obligatoires: irlandais, anglais, mathématiques, éducatif civique, sociale et politique Autres matières: grec, arts, gestion, langues, sciences, etc.
	Évaluation formative / diagnostique	Fin de 1re classe ou début de 2e (6-7a); fin de 4e ou début de 5e (10-11a)	Lecture anglaise et mathématiques
	Pilotage des établissements ou du système éducatif	- 2e classe (4e année enseignement primaire) - 6e classe (8e et dernière année primaire) sur échantillon	NAER: lecture anglaise NAMA: mathématiques
Pays-Bas	Évaluation certificative	Dernière année de primaire (12a)	CITO - Eindloets Basiconderwijs (test de fin de l'enseignement primaire) Langue d'enseignement, arithmétique/mathématiques, capacité d'apprentissage et attitude face au monde extérieur (facultatif)
Suisse alémanique	Évaluation certificative	Zurich: fin de 3e et de 6e	Mathématiques et allemand
	Évaluation formative / diagnostique	- Dans la plupart des cantons, de la 3e à la 9e - 8e et 9e	- Mathématiques et allemand (par ex. Klassencockpit) - Stellwerk: allemand, anglais, français, mathématiques, nature et technique
	Pilotage des établissements ou du système éducatif	- Bâle: fin de 9e - Argovie: en 5e	- Mathématiques et allemand - Mathématiques, allemand et résolution de problèmes
Canada - Ontario	Évaluation certificative	Ontario Secondary School Literacy Test (OSSLT) (prérequis pour le Ontario high school diploma (10e année) si échec Ontario Secondary School Literacy Course à la place du test	Littératie
	Pilotage des établissements ou du système éducatif	- 3e année primaire - 6e année - 9e année	- Lecture, écriture et mathématiques - Junior: lecture, écriture et mathématiques - Mathématiques

en fin de cycle (fin de primaire, fin de scolarité obligatoire) ou parfois au début. Elles peuvent concerner toutes les matières, mais il s'agit le plus souvent des matières principales.

Les évaluations de type diagnostique et/ou formative peuvent avoir lieu à des moments variés des cycles (2e, 3e, 4e, 5e années primaires et à n'importe quel moment du secondaire I). Les disciplines concernées sont plus ou moins les mêmes. Il semblerait que le moment est davantage déterminant que les matières évaluées pour différencier ces trois types d'évaluation.

Le **Tableau 6** permet de comparer pour l'évaluation certificative, l'évaluation de système et l'évaluation formative / diagnostique, les degrés ou niveaux d'enseignement concernés et les domaines. En ce qui concerne la certification, la plupart des pays ou régions ne testent que dans un seul niveau d'enseignement (primaire ou secondaire I). Seuls quelques pays testent les deux niveaux. La majorité des évaluations portent sur les langues ou les mathématiques et moins fréquemment sur les sciences ou les autres domaines. Pour l'évaluation formative/diagnostique, on observe une plus grande diversité des degrés et des niveaux concernés. Les épreuves sont plus nombreuses mais elles sont plus souvent centrées sur les mathématiques et les langues. L'évaluation de système est pratiquée par moins de pays ou de régions. Elle concerne plus souvent le primaire que le secondaire. Par rapport aux deux autres types d'évaluation, l'évaluation de système couvre un plus large spectre de domaines. Si la langue et les mathématiques font quasiment toujours partie de l'évaluation, les sciences et d'autres domaines viennent souvent compléter les domaines évalués.

Élaboration des évaluations

Dans la grande majorité des cas, ce sont des organismes dépendant du ministère de l'éducation qui conçoivent les tests d'évaluation. Dans certains cas, il s'agira plutôt d'une institution publique comme l'université (par exemple au Luxembourg, pour une partie des évaluations) ou un organisme externe privé comme CITO aux Pays-Bas, qui a également fait partie du consortium PISA. Dans d'autres cas, par exemple pour des examens finaux (fin de scolarité obligatoire), les écoles et l'organisme privé se partagent la tâche.

Les concepteurs sont le plus souvent des groupes mixtes composés d'enseignants et d'experts en docimologie; ces groupes peuvent également comprendre des autorités locales (inspecteurs, représentants de l'administration centrale ou de différents services en lien avec l'éducation). Dans deux cas, il s'agit soit d'enseignants seuls, soit d'experts également seuls.

Les évaluations sont construites dans la grande majorité des cas par discipline, très rarement par sous-domaine ou de manière transversale. L'évaluation est constituée le plus souvent d'un panachage de questions à choix multiples et à réponse courte et parfois de questions ouvertes. Dans certains cas, on ne trouve pas de questions ouvertes (par exemple dans les épreuves standardisées visant le monitorage au Luxembourg, dans certains cantons alémaniques ou encore en Italie); dans d'autres, pas de qcm, comme en Communauté française de Belgique pour le test d'enseignement secondaire supérieur (TESS) visant la certification des élèves.

Le format des questions peut aussi dépendre de la discipline évaluée (p. ex. 100% de qcm en langues mais 70% de qcm et 30% de questions à réponse courte en mathématiques). Dans l'ensemble, elles portent sur un mélange d'exercices d'application et de problèmes à résoudre. A part dans un cas, où il est clairement indiqué la proportion de 60% de problèmes à résoudre et 40% d'exercices d'application, les répondants déclarent que cela dépend à la fois de la discipline et du type d'évaluation, voire du public (enseignement général *vs* professionnel).

Dans la plupart des cas, les concepteurs des évaluations procèdent à une répartition équilibrée de questions faciles, moyennes et difficiles, certaines n'utilisant pas de questions moyennes ou de questions difficiles. Comme on le verra plus loin, le degré de difficulté des questions peut être déterminé par des analyses IRT ou de manière plus "qualitative".

Les répondants devaient également donner le nombre de questions prévues pour évaluer un domaine, une notion ou une compétence. Pour la plupart, il a été difficile de répondre précisément à cette question. Pour un domaine, les deux seuls ayant répondu parlent de 50 questions dans un cas et entre 70 et 100 dans l'autre. Les autres évoquent la variabilité en fonction des années et des disciplines mais rappellent également que selon le type d'évaluation, la référence à l'ensemble des objectifs du curriculum mais également le temps imparti doivent être pris en compte.

Pour une notion, la tendance est la même et le nombre est moindre (3-5). Enfin, pour l'évaluation d'une compétence, dans la majorité des cas, il n'y a pas de nombre préétabli, sauf dans un cas où cela varie entre 5 et 20 questions.

Tableau 5. Fonction et fréquence des évaluations externes

Fonction des évaluations	Fréquence	%
Certification élèves	4	33.3
Système + certification	1	8.3
Diagnostique + bilan	1	8.3
Système + bilan + certification	3	25.0
Système + bilan + diagnostique	1	8.3
Système + certification + bilan	1	8.3
Toutes les fonctions	1	8.3
Total	12	100.0

Prétests et analyses des prétests

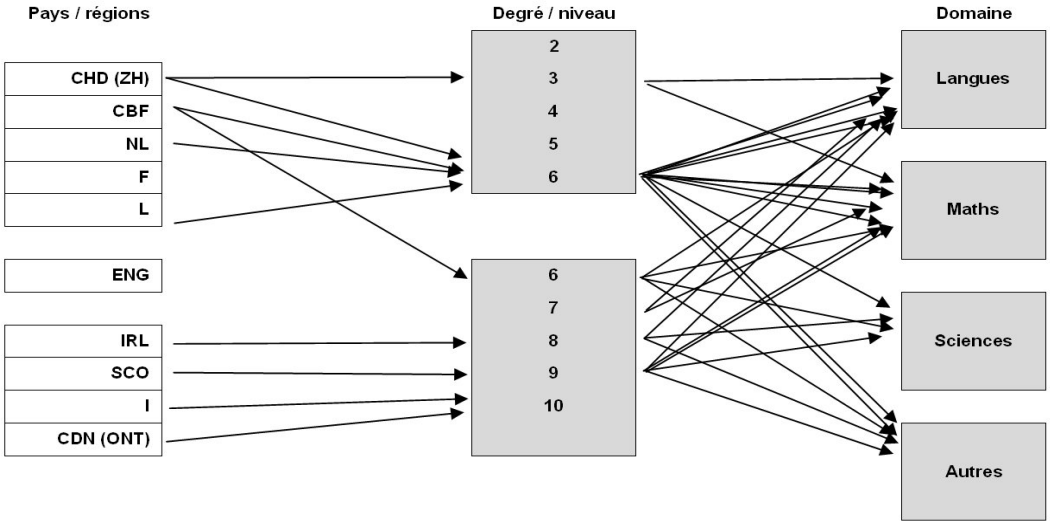
Compte tenu de la situation genevoise où l'on avait pu constater que les prétests n'étaient pas monnaie courante et que les analyses étaient peu pratiquées ou peu détaillées (cf. Soussi et al., 2009), il nous paraissait important de savoir ce qu'il en était dans d'autres contextes.

Dans la grande majorité des cas, les évaluations sont prétestées même si l'on verra qu'elles ne le sont pas forcément dans leur totalité. Le nombre de classes est très variable, allant de 2 à 100 ou 200, dépendant non seulement des méthodes et du type d'évaluation mais également de la taille du pays ou de la région **(Tableau 7)**.

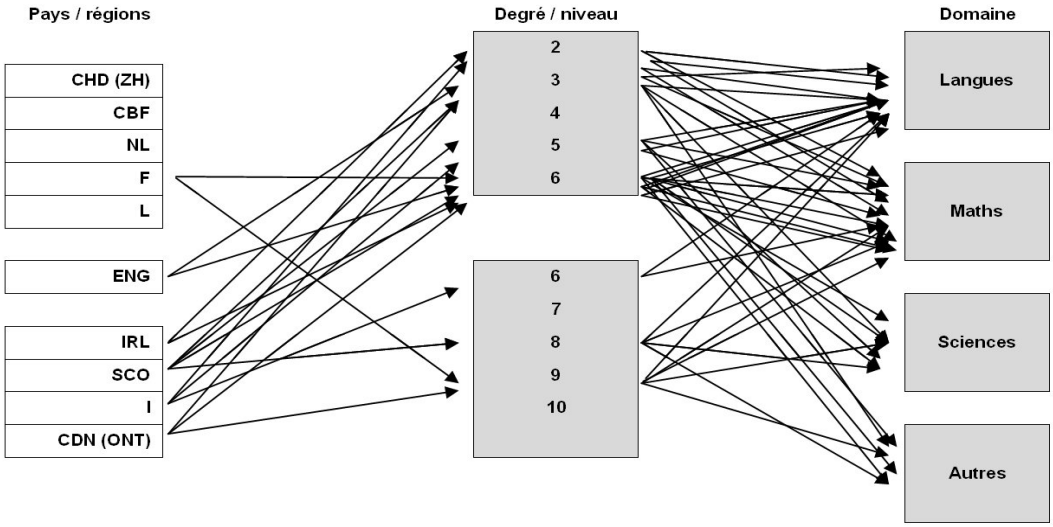
Les analyses réalisées sur les essais peuvent consister en des analyses d'items simples ou passer par des méthodes statistiques plus sophistiquées de type IRT. Ces deux types d'analyses peuvent également être utilisées conjointement; les analyses d'items peuvent également être accompagnées d'analyses par la méthode des juges. Dans un certain nombre de cas, des items supplémentaires ont été prévus dans le prétest, variant de 1/3 à 50-100% **(Tableau 8)**. Le nombre d'items éliminés est très variable d'une évaluation ou d'un pays à l'autre. Ils peuvent aller jusqu'à 50% des items testés.

Tableau 6. Moments du cursus et domaines évalués selon la fonction de l'évaluation

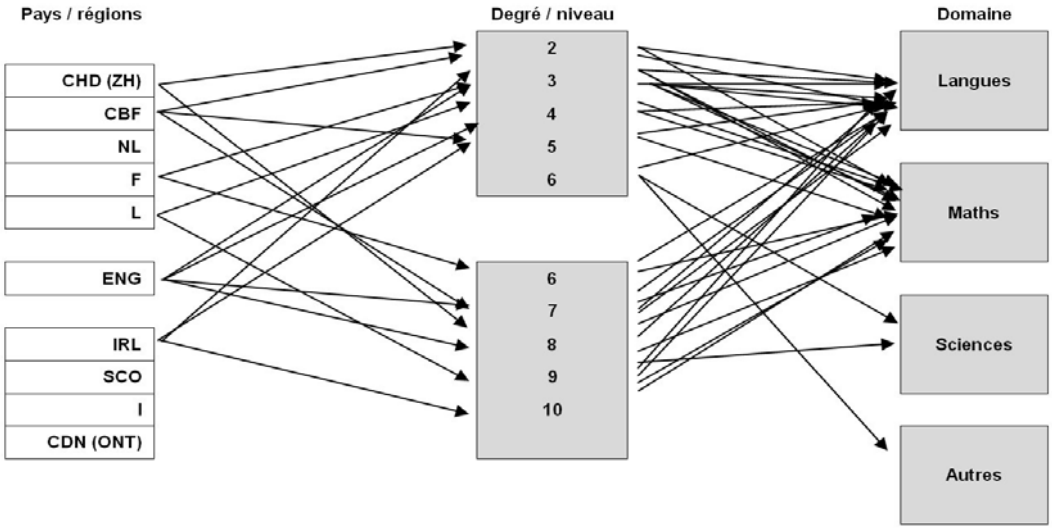
Certification



Monitorage système



Formative / diagnostique



Passation et correction

Dans la majorité des cas, la passation est assurée par les enseignants de l'école – qu'il s'agisse de ceux des élèves ou d'autres – et parfois également par des personnes externes. Pour la correction, la situation est un peu différente et dépend du type d'évaluation. Dans certains cas d'évaluation du système ou de monitorage, ce sont des personnes externes qui sont chargées de la correction.

Le plus souvent, la passation ne fait pas l'objet d'une formation mais un guide est fourni aux enseignants ou aux autres personnes qui font passer le test. Ce guide comporte en général les consignes de passation ou les objectifs des questions ou encore les deux. La correction fait un peu plus souvent l'objet d'une formation. Dans les deux tiers des cas, un guide de correction est également fourni, qui contient parfois en outre les objectifs des questions.

Analyse et diffusion des résultats

L'analyse des résultats fournit avant tout les scores globaux des élèves. A cela peuvent s'ajouter une analyse par domaine ou par item, voire les deux. L'analyse par item va souvent de pair avec une analyse IRT. Les scores globaux peuvent aussi être accompagnés d'analyse prenant en compte les caractéristiques des élèves. Pour ce qui concerne les barèmes, dans la grande majorité des cas, ils sont calculés a priori par objectif ou compétence. Ils font quelquefois l'objet d'un ajustement a posteriori en fonction de la distribution des résultats.

La diffusion des résultats est souvent générale, c'est-à-dire pour le public. Dans quelques cas plus isolés, ils ne sont diffusés qu'aux directions d'école et aux enseignants, ou aux enseignants et aux élèves, tout dépend du type d'évaluation. Le format des résultats diffusés (rapport ou document plus succinct) dépend à la fois du public et du type d'évaluation. Aux Pays-Bas et en Ontario par exemple, on trouve des rapports sous différentes formes directement téléchargeables.

Régulation des évaluations

Ce dernier point nous paraissait important à investiguer. Souvent, les évaluations sont analysées et les résultats diffusés dans l'urgence et le temps manque pour tenir compte des éventuels problèmes survenus (cf. la situation genevoise décrite dans l'étude du SRED, 2009). Dans certains cas, ce ne sont pas les mêmes personnes qui conçoivent les tests, ce qui rend la régulation difficile.

Les réponses fournies par nos répondants sont variées, allant d'une absence de régulation à des efforts plus ou moins conséquents: par exemple en CFB, l'élaboration des épreuves de l'année suivante peut commencer par une analyse de celle de l'année d'avant; ou un questionnaire bilan est envoyé aux directeurs et aux enseignants afin de récolter leurs impressions sur le dispositif complet de l'évaluation (passation, correction, analyse et exploitation des résultats ainsi que pistes didactiques); ou encore, comme au Luxembourg, on essaie de prendre progressivement en considération les remarques et propositions des acteurs du terrain.

Tableau 7. Taille de l'échantillon faisant l'objet d'un prétest

Taille de l'échantillon (nb. de classes ou d'élèves)	Fréquence	%
Non-réponses	2	16.7
Une dizaine de classes	1	8.3
100 à 200 classes représentatives du niveau considéré (taille des classes, appartenance à un type d'école public, privé, éducation prioritaire, rural/urbain)	1	8.3
2 classes par version de texte (càd 2x3 branches x3 filières)	1	8.3
5 classes sélectionnées pour constituer 1 échantillon le plus hétérogène possible (en fonction de la zone géographique, du réseau d'enseignement et du statut socioéconomique du public)	1	8.3
Chaque livret est testé sur 2 classes de 6e	1	8.3
Environ 300 réponses par item : échantillon représentatif en fonction du genre, des compétences, etc. (quelques items du test, pas son ensemble)	1	8.3
Les items sont prétestés dans quelques classes en enfouissant quelques items du prétest dans une évaluation opérationnelle (cf. rapports techniques disponible sur le site EQAO)	1	8.3
Minimum 150 élèves par épreuve, provenant si possible de 10 écoles différentes et contrastées (en fonction de leur zone géographique, leur réseau d'enseignement, de la filière de formation choisie, général ou prof.)	1	8.3
La taille de l'échantillon dépend du type de test	1	8.3
Pour chaque test, il y a un échantillon national de 400 élèves (20 classes)	1	8.3
Total	12	100.0

Synthèse et discussion

Comme on a pu l'observer dans ce qui précède, les situations sont variées d'un pays ou d'une région à l'autre, certains privilégiant davantage l'une ou l'autre forme d'évaluation externe. Dans certains pays, on a une longue tradition d'évaluation externe et on utilise des méthodes scientifiques sophistiquées en utilisant l'analyse de réponse à l'item pour analyser les prétests et construire les évaluations, en contrôlant la fidélité et la validité de l'épreuve, etc. Dans d'autres cas, les méthodes sont plus qualitatives. Ces différentes méthodes ne préjugent pas forcément de la qualité des épreuves.

Dans la plupart des pays, les autorités scolaires ont mis en place des évaluations externes pour juger de l'efficacité de l'enseignement. Selon le type d'évaluation choisi, les effets sont toutefois différents : certaines visent la comparabilité entre établissements, d'autres au travers de la certification permettent de vérifier l'atteinte des objectifs et d'harmoniser les pratiques, d'autres enfin peuvent contribuer à la régulation de l'enseignement. Les évaluations externes peuvent être regroupées selon les différentes fonctions qu'elles assurent: monitorage du système, certification des élèves ou évaluation formative ou diagnostique des élèves. Au niveau des pays ou des régions, ce sont les deux premières fonctions que l'on retrouve le plus souvent; la dernière, l'évaluation formative / diagnostique est plus rare, étant sans doute davantage pratiquée au niveau interne.

Ces évaluations portent surtout sur la langue d'enseignement et les mathématiques. Au niveau de la Suisse romande, on trouve essentiellement des évaluations externes certificatives ou formatives / diagnostiques. Cependant, le panorama est assez différent selon les cantons. A part Genève et Vaud (dans une moindre mesure) qui n'utilisent que des évaluations certificatives de façon assez intensive, la plupart des cantons recourent aux deux types d'évaluation, l'évaluation de type certificatif intervenant essentiellement au moment de la transition entre les degrés d'enseignement (8P et 11e).

En Suisse romande, les évaluations externes n'ont pas pour but explicite le monitorage du système. Le sondage réalisé auprès de quelques pays ou régions met en évidence que les évaluations certificatives portent le plus souvent soit sur le primaire soit sur le secondaire I, selon les pays et régions, et que l'évaluation formative / diagnostique est plus diversifiée en fonction des degrés et niveaux d'enseignement.

Dans la majorité des cas, les épreuves sont conçues par des groupes mixtes d'enseignants et d'experts en docimologie d'organismes dépendant des ministè-

res de l'éducation. Elles sont élaborées selon une logique disciplinaire. La plupart des épreuves sont prétestées, cependant ces prétests peuvent être de nature et d'envergure très différentes. Ils peuvent être réalisés de façon relativement qualitative avec peu d'élèves concernés ou au contraire être d'une envergure plus large.

On observe également que certains de ces prétests testent l'ensemble de l'épreuve alors que dans d'autres cas, on préteste des items ou des parties d'épreuves. Dans quasiment tous les cas, il est prévu un nombre d'items plus élevé que nécessaire (de 10% à près de 50% d'items supplémentaires).

Par ailleurs, l'administration des tests et la correction sont réalisées par des enseignants. Cependant, ce ne sont pas forcément les enseignants des élèves; les corrections peuvent aussi être effectuées par d'autres enseignants que ceux des élèves. Ce n'est toutefois que pour les évaluations de système que l'on recourt parfois à des personnes externes aux écoles. Quant à la diffusion des résultats, ils sont le plus souvent communiqués sous forme de résultats globaux et publics.

La situation genevoise

Comment se situe Genève dans ce contexte? On notera tout d'abord que comparativement aux autres cantons, régions et pays étudiés, Genève se caractérise par un recours important, au niveau de la scolarité obligatoire, à une forme d'évaluation externe visant essentiellement la certification des élèves dans un nombre croissant de disciplines entre le primaire et le secondaire I, avec des modalités différentes d'élaboration des épreuves en fonction des degrés d'enseignement (cf. Soussi, Guilley, Guignard et Nidegger, 2009).

Par rapport à ce qui ce se fait ailleurs, on retiendra que Genève ne recourt pas de façon systématique aux prétests, que la construction des épreuves est réalisée surtout par des enseignants ou des formateurs sous la supervision de coordinateurs avec un regard docimologique.

Des équipes pluridisciplinaires avec des experts en docimologie pourraient contribuer à améliorer la validité et la fidélité de la mesure en dosant par exemple le degré de difficulté des items ou en vérifiant le nombre d'items requis pour l'évaluation d'une compétence.

Tableau 8. Nombre d'items supplémentaires prévus pour le prétest

Pourcentage ou nombre d'items supplémentaires prévus	Fréquence	%
Non-réponses	1	8.3
Une quinzaine	1	8.3
Environ un tiers	4	33.3
50 à 100%	1	8.3
100%	1	8.3
Dépend des besoins et du moment du cursus	1	8.3
Pas de proportion préétablie	2	16.7
Ne sait pas	1	8.3
Total	12	100.0

La mise en place de prétests systématiques et leur analyse seraient également un atout supplémentaire pour les épreuves genevoises qui ont déjà une longue tradition. Les conditions de passation et de correction pourraient également gagner en efficacité en croisant par exemple les classes.

De plus, si les résultats diffusés, comme dans beaucoup de pays, sont des scores globaux, on peut regretter que les données disponibles soient si peu exploitées. On a pu également constater, à Genève comme ailleurs, le manque de régulation d'une année à l'autre faute de temps ou de ressources.

Comme nous l'annoncions déjà en 2009, la grande question sera de déterminer la place des évaluations cantonales genevoises dans le contexte romand (Eprocom) et national (tests de référence HarmoS). Si pour l'instant les épreuves genevoises sont plutôt de type certificatif et que les tests de référence HarmoS auront plutôt pour objectif le monitorage et le pilotage du système, quel rôle les épreuves romandes devraient-elles jouer ? Seront-elles certificatives, diagnostiques ou encore bilan (mesure de l'atteinte des objectifs du PER) ?

La question est particulièrement complexe dans la mesure où les épreuves genevoises, si elles continuent d'exister, porteront comme les épreuves romandes sur le PER. Il s'agira pour les trois types d'évaluations externes de se positionner au moins sur trois critères: la fonction, les matières évaluées et bien sûr les moments du cursus, afin d'éviter qu'une trop grande partie du temps scolaire ne soit consacrée à l'évaluation. Il est malheureusement encore trop tôt pour le dire. ■

Notes

[1] Les pays qui ont fait l'objet de l'enquête sont indiqués en italiques dans le texte.

[2] Ce tableau est tiré de Marc, V. & Wirthner, M. (2012). *Épreuves romandes communes : de l'analyse des épreuves cantonales à un modèle d'évaluation adapté au PER : rapport final du projet EpRoCom*. Neuchâtel : IRDP (12.1).

[3] Ce sondage a été conçu avec la collaboration de E. Guilley et N. Guignard (SRED).

Références

- Eurydice (2009). *Les évaluations standardisées des élèves en Europe: objectifs, organisation et utilisation des résultats*. Bruxelles: Eurydice.
- Mons, N. (2009). "Les effets théoriques et réels de l'évaluation standardisée". Complément à l'étude *Les évaluations standardisées des élèves en Europe: objectifs, organisation et utilisation des résultats.* Eurydice.
- Ntamakiliro, L. & Tessaro, W. (2010). *Les évaluations externes en Suisse romande: enjeux, pratiques et effets.* Communication lors de la rencontre de la section suisse de ADMEE-Europe portant sur le thème "Épreuves cantonales: quelles pratiques pour quels enjeux ?" Martigny.
- Soussi, A., Guilley, E., Guignard, N., Nidegger, Ch. (2009). *Évaluation des acquis des élèves à l'école obligatoire*. Genève: SRED.
- Weiss, J. (2002). "L'évaluation externe dans un concept général et cohérent d'évaluation". *Actes du 15e colloque international de l'ADMEE-Europe et Congrès annuel de la SSRE* (pp. 2-5). Lausanne: ISPFP.

Version électronique de cette note : http://www.ge.ch/recherche-education/doc/publications/notesinfo/notes-sred-54.pdf

Informations complémentaires : anne.soussi@etat.ge.ch, 022 546 71 39 - christian.nidegger@etat.ge.ch, 022 546 71 19

Edition : narain.jagasia@etat.ge.ch, 022 546 71 14

Département de l'instruction publique, de la culture et du sport
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